On the Influence of Alcohol Intoxication on Speaker Recognition

Jürgen T. Geiger[1], Boxin Zhang[1], Björn Schuller[2,1], and Gerhard Rigoll[1]

[1]*Institute for Human-Machine Communication, Technische Universität München, Munich, Germany*

[2]*Department of Computing, Imperial College London, London, UK*

Correspondence should be addressed to Jürgen T. Geiger (geiger@tum.de)

ABSTRACT

In this paper we present a study of the influence of alcohol intoxication on an automatic speaker verification system. It is widely known that alcohol intoxication affects one's speech in many ways, but it remains to be studied how well a system can recognise a person affected by alcohol intoxication. Using the Alcohol Language Corpus as speech database and a text-independent GMM-UBM speaker verification system, we perform experiments to analyse the effects of alcohol intoxication in detail. In different experimental setups, using recordings in either sober or alcoholised condition for speaker enrolment or testing, the influence of intoxication on the error rate of the speaker verification system is investigated. Compared to the baseline experiment without alcohol intoxication, the results indicate a generally negative influence of alcohol intoxication on the verification system. This influence is larger for female speakers compared to males.

1. INTRODUCTION

It is known that alcohol intoxication as well as some other factors such as health state, stress and fatigue affect one's speech [1, 2]. This holds in, for instance, the content of the speech and the physical acoustic signal. Thereby, the speech can be affected in two ways: degradation of the neurological function of human and degradation of the motor control ability. There exist several studies related to the effects of alcohol on speech. A review of these effects is given in [3]. The study presented in [4] confirms that the fundamental frequency and its range are raised with increasing intoxication. In [5], one of the first studies trying to detect intoxication from speech is presented. In [6], changes in glottal pulse parameters were used to detect alcohol intoxication. A relatively large speech corpus containing recordings under alcoholised and sober condition is presented in [7]: the Alcohol Language Corpus (ALC). It contains alcoholised speech recordings taken in an automotive environment, and enables reliable studies on the topic of intoxication influence on automatic speech processing systems. In the Interspeech 2011 Speaker State Challenge [8, 9], the ALC data were used as a benchmark to test systems for intoxication recognition. Following this challenge, the effect of alcohol intoxication on human speech was studied in more detail. For example, in [10], results showed that human listeners use prosodic information to detect alcohol intoxication. In [11], alcohol intoxication was regarded as a different accent of a speaker, and phonetic, phonotactic and prosodic cues were used to detect intoxication. Furthermore, text-based features can be an indicator for alcohol intoxication [12].

Although there have been several studies investigating the influence of intoxication on the characteristics of human speech, there are no studies, to the best of our knowledge, on the direct influence of alcohol intoxication on speech processing systems like automatic speech or speaker recognition. Thus, it is interesting to investigate the influence of alcohol intoxication on a speaker recognition system and to discover the performance of a speaker recognition system using the alcoholised speech from the ALC corpus. Generally, it is expected that using alcoholised speech degrades the system's ability to recognise speakers. There are several possible applications of speaker recognition systems where alcohol intoxication may influence the system performance, such as home access systems, telephone hotlines or access systems to public events with lots of drunk people. An interesting question is to what degree the performance is influenced by alcohol intoxication. Also, to what extent

is a speaker harder to recognise if he/she is intoxicated and the speaker model was built using 'sober' data? Additionally, it seems interesting to know if, in a speaker verification system, using alcoholised or sober impostors influences the system performance. Furthermore, does it help to use alcoholised recordings as training data? Finally, when alcoholised data are used for target model creation, how well can the system recognise sober speakers?

In this paper, we thus analyse the effects of alcohol intoxication on a speaker verification system. Using recordings in sober and alcoholised condition from the same set of speakers, the influence of intoxication can directly be measured in terms of recognition performance. A state-of-the-art speaker verification system applying the Gaussian Mixture Model (GMM) approach with a Universal Background Model (UBM) and Linear Frequency Cepstral Coefficients (LFCCs) is used for experiments with the ALC data for training and evaluation. Several experiments are conducted, using either sober or alcoholised speech material for model training and for the true speaker and impostor trials.

The rest of the paper is organised as follows: In Section 2, the theoretical concepts and processes of automatic speaker recognition and the employed system are introduced. Section 3 presents the experimental methods and shows the results of the experiments. Finally, in Section 4, some conclusions are given.

2. EMPLOYED SPEAKER VERIFICATION SYSTEM

Speaker verification refers to a task where an unknown speaker claims to be a specific identity and the system should make the decision whether to confirm or deny the claimed identity. Speaker verification is also known as speaker authentication; it is a voice match between the speaker's speech and the true identity. A speaker verification system is composed of two distinct processes: an enrolment process and a recognition process. In the enrolment process, a target speaker model is trained after a model adaption with the background model (or world model). The feature parameters for training statistical models, including both the world model and target model, are extracted from the speech signal in the first place. In the recognition process, the feature vectors of an unknown speaker are compared with the speaker model, giving a score of similarity between them both.

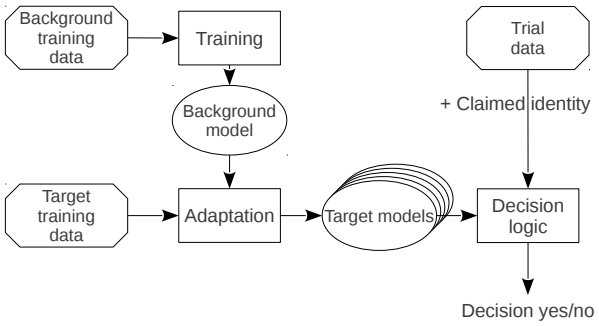

Fig. 1: Functionality of the employed speaker verification system

The decision module makes a final decision based on the similarity score and a threshold.

2.1. GMM-UBM System

GMM-UBM is a well-established approach in speaker recognition [13]. The functionality of a GMM-UBM system, as it is applied in this work, is depicted in Fig. 1. In GMM-based text-independent speaker recognition, a world model or universal background model (UBM) is trained with the Expectation Maximisation (EM) algorithm for usually five to ten iterations from a large set of speakers, resulting in a single model to represent the speaker-independent distribution of features. The UBM is then used as a common reference for training of the other speaker models. When a new speaker is enrolled to the system, the UBM parameters are adapted to the feature distribution of the new speaker. It is to be noticed that the world model is trained with the Maximum Likelihood Parameter Estimation (ML), where the criterion here is to maximise the likelihood of the data that are computed by the model. On the contrary, the target speaker model training is realised by maximising the a posteriori probability that the claimed identity is true. This method is called maximum a posteriori (MAP) adaptation. The approach of training a background model with the ML method and then adapting to the target model training using the MAP criterion is referred to as GMM-UBM [14].

We applied the following system parameters: For world model building, 1 024 Gaussian distributions were used. Preliminary experiments were performed where the number of mixture components was altered. Results showed the general trend that doubling the number of

components (starting from 64 distributions) decreased the equal error rate (EER), at the cost of higher computational complexity. Therefore, 1 024 distributions were chosen. The number of iterations for training the world model was set to 6, during which 10 % of the frames were taken for each iteration. For target speaker model building, five iterations were used, in which 100 % frames were used. Smaller numbers of iterations led to similar results. However, when using less than 2 iterations, the EER increases drastically. The MAP regulation mean factor was set to 7, which proved to be a good choice. Finally, in the step for computing both the true speaker and impostor scores, we computed only the first 10 top Gaussian distributions. Our speaker recognition system is built based on the open-source platforms ALIZE and LIA_RAL [15].

2.2. Feature Extraction

The feature extraction was carried out with 16 Linear Frequency Cepstral Coefficients (LFCCs). Compared to conventional Mel Frequency Cepstral Coefficients (MFCCs), LFCCs showed to perform better in speaker recognition tasks especially for female voices, due to the shorter vocal tract in females and the resulting higher formant frequencies. LFCCs can better capture the spectral characteristic in the high frequency region [16]. By adding the first order derivatives, log energy and delta energy, we obtain a feature vector of the size of 34 dimensions. Features are extracted every 10 ms from Hamming windows with a length of 20 ms. Feature extraction was performed with the SPro toolkit[1].

3. EXPERIMENTS

This section describes the employed speech corpus, experimental setup and results.

3.1. The Alcohol Language Corpus

Experiments are performed using the ALC data for training and testing. ALC is a publicly available speech database. It contains recordings from 162 german speakers (78 female and 84 male) from 5 locations in Germany, ages 21 – 75 years (mean age 31 years). It covers a variety of speech styles and speaker types. For each speaker, recordings containing read, spontaneous and command & control speech in various forms were taken. This means that the recordings contain a certain amount of realistic conversational speech. Thus, further research can benefit from the diversity of the database and produce realistic results. To build the corpus, the speakers underwent a systematic intoxication test. Each speaker chose voluntarily her/his blood alcohol concentration (BAC). In the corpus, the BAC values range from 0.28 to 1.75 per mill[2]. The required amount of alcohol for individuals was estimated with the Watson- and Widmark formula [17]. Speakers had to take a break of twenty minutes after having consumed the estimated amount of alcohol. Then the speaker underwent a breath alcohol concentration (BRAC) test and a blood sample test. A 15-minute ALC speech test for each individual speaker followed immediately, so that the changes due to fatigue and some other factors were reduced. The second recording was taken after two weeks; it lasted for 30 minutes under sober condition, which means without alcohol intoxication.

3.2. Experimental Setup

The partition of data sets for the experiments is shown in Table 1. The UBM is created using 20 recordings each from 51 speakers. For each speaker, the recordings were randomly selected. For target model training, true speaker trials, and impostor trials, the same 44 speakers were selected, using 30 utterances per speaker (summing up to 4.0 hours of speech data for each set). The speakers that are used for UBM training and for the trials are selected from the training and development set, respectively, as defined in the Interspeech 2011 Speaker State Challenge. In the case of intoxicated recordings, the BAC values from those 44 speakers are in the range from 0.3 to 1.7 per mill, with a variance of 0.09 per mill. 30 recordings are used to create the target models for each speaker. For the true speaker trials, each feature vector of a speech segment was tested against the only true target speaker model. On the other hand, for the impostor trails, we tested those feature vectors of each speech segment against 10 randomly picked faked identities to obtain a score for impostors. Thus, each experiment is composed of 1 320 true speaker trials and 13 200 impostor trials, summing up to 14 520 trials in total.

Considering that in a real world application most of the speech recordings made by subjects are without alcohol influence, we built the UBM in all our experiments with sober data only, though it would also be interesting to

[1] http://www.irisa.fr/metiss/guig/spro/

[2] Per mill BAC by volume (standard in most central and eastern European countries)

	speakers	recordings per spk.
train UBM	51	20
train target	44	30
true speaker trial	44	30
impostor trial	44	30

Table 1: Data sets used for the experimental setup

design experiments in which we train the UBM using mixed data or only alcoholised. For target model training and true speaker trials, we tested both settings, i. e., using sober or alcoholised recordings. Using alcoholised recordings for target model training can be considered a variant of multi-condition training, since the UBM is trained with sober data. We examined the influence of two kinds of impostors: sober and alcoholised, in order to find out if there was some difference that impostor tests could make to the system performance in verifying speakers under alcohol intoxication.

There are two types of errors in a speaker verification system: false rejection (FR) and false acceptance (FA) [18]. The threshold θ is applied in the decision making process. With a higher threshold, more identity claims would be rejected, resulting in more false rejections but fewer false acceptances. On the contrary, when the threshold θ is set to some lower value, more claims would be accepted, so we get more false acceptance but few false rejections. Thus, setting the threshold θ is a trade-off between two types of errors. The probabilities for false rejections and false acceptances, P_{fr} and P_{fa} are obtained during experiments, specifically in the test phase by computing the numbers of errors of each type. The couple error rates P_{fr} and P_{fa} are both functions of the threshold. Thus, we can plot P_{fa} as a function of P_{fr}. This curve is known as the detection error trade-off (DET) curve. It has been a standard method to represent the system operating characteristic [19]. In a speaker recognition system where the true and impostor speaker scores can be assumed as being Gaussian distributed and having the same variance, the DET curve would be linear and have a slope of -1. In practice, the score distributions are very close to Gaussians though not exactly, which ensures the capability of representing the system performance. The DET curve is a more linear graph, compared to the traditional Receiver Operating Characteristic (ROC) curve, so it is more intuitive to read [20, 21]. Another metric used to evaluate the system performance is the EER. It corresponds to the intersection of the DET curve with the bisector line and indicates the operating point in the DET curve where $P_{fr} = P_{fa}$. An issue with EER is that it indicates an arbitrary decision threshold, showing the overall performance of the system [22].

	UBM	tr.tgt	true trial	impostor	EER[%]
Exp. 1	S	S	S	S	8.1
Exp. 2	S	S	A	S	12.9
Exp. 3	S	S	A	A	12.3
Exp. 4	S	A	S	S	10.9
Exp. 5	S	A	A	S	8.1
Exp. 6	S	A	A	A	7.9

Table 2: Alcohol intoxication setup (S: sober, A: alcoholised) for UBM training, target training (tr.tgt), true speaker trials and impostor trials, and results in terms of EER.

3.3. Results

This section shows the relevant results of the experiments that were performed as described above. The performance was evaluated by plotting the DET curve and computing the EER. Various experimental setups with different configurations of sober or intoxicated data and corresponding results are shown in Table 2.

As the target models are trained separately with sober (Exp. 1-3) and alcoholised (Exp. 4-6) data, we compare the results separately within these two setups.

In the first part of the experiments, the target model is trained using only sober data. Fig. 2 illustrates the DET curves for Experiments Exp. 1–3. Exp. 1 (solid line) can be considered the baseline experiment, since only sober recordings are used for training and testing. The EER of the baseline system is 8.1 %. Using alcoholised data for both true and impostor trials (Exp. 3), results in an EER of 12.3 %. The difference in EER between these two setups is 4.2 %, which means that, when having sober target models and test with alcoholised data, the system's EER is generally 4.2 % worse than with only sober data. For Exp. 2, true trials use alcoholised data and imposter trials are done with sober recordings. In this case, the EER is increased to 12.9 %.

Next, let us take a closer look at Exp. 4, Exp. 5, and Exp. 6 in Table 2. These experiments are conducted

under the condition that the target speaker models are trained using alcoholised data. The corresponding DET curves are shown in Fig. 3. The solid curve is generated with the setup for Exp. 4 and has an EER of 10.9 %, the dashed line represents Exp. 5 (EER 8.1 %), while the dashed-dotted one is the result of Exp. 6 and has an EER of 7.9 %. Comparing these three curves, we can see that, when we train the target with alcoholised data, then test with only sober data, the EER is 3% worse, compared to the 'matched-condition' experiment, when alcoholised data are used for testing.

Finally, we observe the impostor tests, which were taken using sober or alcoholised data in different experiments in Table 2. For instance, we compare Exp. 2 and Exp. 3. They show the EERs of 12.9 % and 12.3 %. The comparison yields that slight differences exist between these two setups of impostor trials. The increase of EER from 12.3 % to 12.9 % is attributed to a higher P_{fa}. Since the target models are trained with sober data, using sober impostors increases the number of false alarms, compared to when using alcoholised impostors. A similar observation can be made by comparing Exp. 5 and Exp. 6, which have EERs of 8.1 %, 7.9 %, respectively; yet, this difference is not statistically significant. However, even when target models are trained with alcoholised data, using sober impostors increases P_{fa}. This could lead to the assumption (even if the difference is not significant) that the target models are not fully adapted to the alcoholised case.

It is to be noticed that all three results in Fig. 3 are better when comparing to the results in Fig. 2. One can see the difference when comparing the baseline system (8.1 %) against Exp. 6 (7.9 %), or Exp. 3 (12.3 %) against Exp. 4 (10.9 %). These interesting results may indicate that (under the assumption that $p(A) = p(S)$, which, however, is generally not true), it might generally be better to train target models using alcoholised data rather than sober data to obtain a better system performance. This is a result of the fact that in Exp. 4-6, the target models are adapted from a sober background model, using alcoholised data, which can be regarded as a sort of multi-condition training.

The experimental results are then further analysed for the differences between male and female speakers. We compared the results of the experiment with sober data only (Exp. 1) with the results obtained by using sober models and alcoholised trial data (Exp. 3), separately for male and female speakers. These results are shown in Table 3.

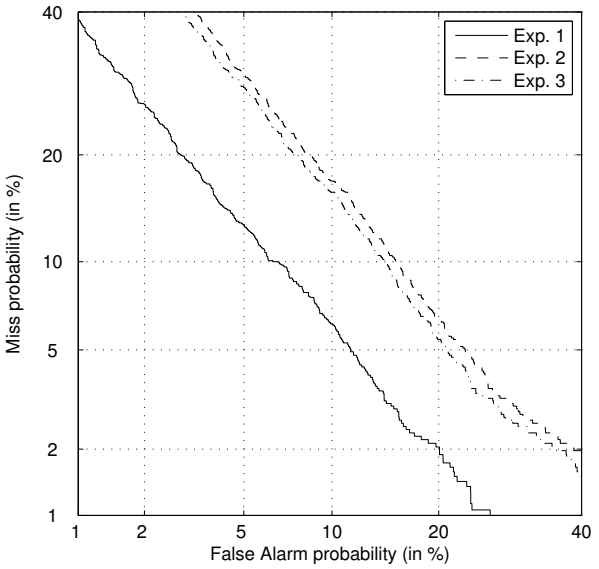

Fig. 2: Speaker verification DET curves when target models are trained using sober data. Solid line: sober trials (Exp. 1, 8.1 % EER); dashed line: alcoholised true speaker trials and sober impostor trials (Exp. 2, 12.9 % EER); dashed-dotted line: alcoholised trials (Exp. 3, 12.3 % EER). For explanation of experimental configurations, cf. Table 2.

For male speakers, using alcoholised recordings for testing increased the EER by 3.5 %, while, for female speakers, the EER increased by 4.8 %. Thus, in our experiment, alcohol intoxication has a larger effect on female speakers compared to male speakers. However, in order to be able to draw general conclusions, more experiments need to be performed, since the sample size is not large enough to produce statistically significant results in our case.

4. CONCLUSIONS

In this work, an automatic speaker verification system based on GMM-UBM approach is used for analysing the influence of speaker alcohol intoxication on the system performance. A set of experiments was conducted with different configurations of using alcoholised or sober data for training or testing. The results were shown and discussed in detail, generally indicating a negative influence of alcohol intoxication on the speaker verification system in the case of mismatched conditions. We

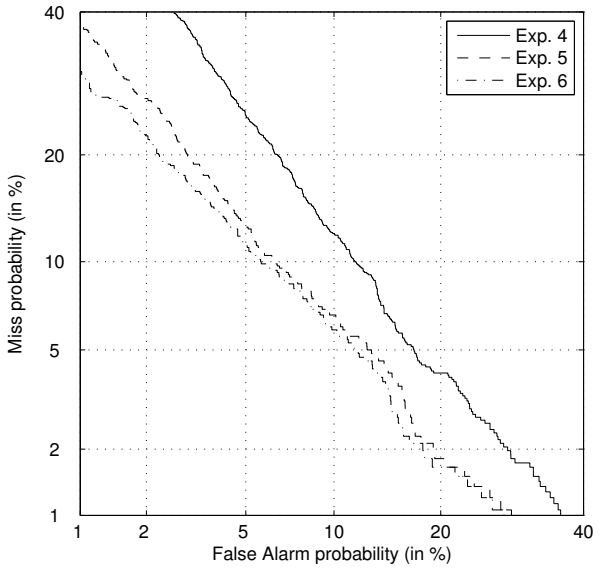

Fig. 3: Speaker verification DET curves when target models are trained using alcoholised data. Solid line: sober trials (Exp. 4, 10.9% EER); dashed line: alcoholised true speaker trials and sober impostor trials (Exp. 5, 8.1 % EER) ; dashed-dotted line: alcoholised trials (Exp. 6, 7.9% EER). For explanation of experimental configurations, cf. Table 2.

also found that training target speaker models with alcoholised data always gives a better result than using sober data. This can be attributed to the usage of sober data for background model training, which, in combination with alcoholised data for target model training, can be considered as multi-condition training. An analysis of the results showed that alcohol intoxication had a larger influence on females compared to males.

Future work could explore the system performance based on different degrees of alcohol intoxication or different speech styles, to find out how verification rates are affected gradually by varying conditions. Approaches for multi-task learning could be used to directly detect intoxication and recognise the speaker. Here we playn to adapt and use our methods developped for detecting overlapping speech: explpoiting linguistic information [23] or using neural networks for classification [24]. Another idea is to apply techniques to optimise the intoxicated case, such as techniques for session variability or channel compensation, e. g., Joint Factor Analysis [25], in which the two states of our system (sober or alcoholised)

EER[%]	male	female
Exp. 1	8.8	7.1
Exp. 3	12.3	11.9
difference	+3.5	+4.8

Table 3: Gender-dependent results, comparing the EER in the case of taking only sober data (Exp. 1) or taking sober data for training and alcoholised data for testing (Exp. 3).

could be regarded as different 'telephone channels'. Furthermore, it would be interesting if directly addressing the effects of alcohol intoxication (e. g., increased F0 or range of F0) and trying to reverse them can help to increase the robustness of a speaker recognition system to alcohol intoxication.

5. REFERENCES

[1] J. HL Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 1, pp. 151–173, 1996.

[2] H. JM Steeneken and J. HL Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. ICASSP*, Orlando, FL, USA, 1999, pp. 2079–2082.

[3] S. B. Chin and D. B. Pisoni, *Alcohol and Speech*, Academic Press, 1997.

[4] B. Baumeister, C. Heinrich, and F. Schiel, "The influence of alcoholic intoxication on the fundamental frequency of female and male speakers," *The Journal of the Acoustical Society of America*, vol. 132, pp. 442, 2012.

[5] K. Johnson, D.B. Pisoni, and R.H. Bernacki, "Do voice recordings reveal whether a person is intoxicated? A case study," *Phonetica*, vol. 47, no. 3-4, pp. 215–237, 1990.

[6] M. Sigmund and P. Zelinka, "Analysis of voiced speech excitation due to alcohol intoxication," *Information Technology and Control*, vol. 40, no. 2, pp. 143–150, 2011.

[7] F. Schiel, C. Heinrich, S. Barfüßer, and T. Gilg, "ALC: Alcohol language corpus," in *Proc. LREC*, Marrakesch, Marokko, 2008.

[8] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3201–3204.

[9] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker statesa review on intoxication, sleepiness and the first challenge," *Computer Speech & Language, Special Issue on Broadening the View on Speaker Analysis*, 14 pages, 2012.

[10] F. Schiel, "Perception of alcoholic intoxication in speech," in *Proc. of Interspeech*, Florence, Italy, 2011, pp. 3281–3284.

[11] F. Biadsy, W. Y. Wang, A. Rosenberg, and J. Hirschberg, "Intoxication detection using phonetic, phonotactic and prosodic cues," in *Proc. of Interspeech*, Florence, Italy, 2011, pp. 3209–3212.

[12] T. Bocklet, K. Riedhammer, and E. Nöth, "Drink and speak: On the automatic classification of alcohol intoxication by acoustic, prosodic and text-based features," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3213–3216.

[13] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[14] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[15] J.F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2008, vol. 5.

[16] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. ASRU*, Big Island, HI, USA, 2011, pp. 559–564.

[17] P. Watson, I. Watson, and R. Batt, "Prediction of blood alcohol concentrations in human subjects; updating the widmark equation," *Journal of Studies on Alcohol and Drugs*, vol. 42, no. 07, pp. 547, 1981.

[18] D.A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the Switchboard corpus," in *Proc. ICASSP*, Atlanta, GA, USA, 1996, pp. 113–116.

[19] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. of Eurospeech*, Rhodes, Greece, 1997, pp. 1895–1898.

[20] C.E. Metz, "Basic principles of roc analysis," in *Seminars in Nuclear Medicine*. Elsevier, 1978, vol. 8, pp. 283–298.

[21] J.A. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*, Lawrence Erlbaum Associates, Inc, 1996.

[22] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[23] J. Geiger, F. Eyben, N. Evans, B. Schuller, and G. Rigoll, "Using linguistic information to detect overlapping speech," in *Proc. Interspeech*, Lyon, France, 2013, pp. 690–694.

[24] J. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proc. Interspeech*, Lyon, France, 2013, pp. 1668–1672.

[25] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.