Towards intoxicated speech recognition

Zixing Zhang, Felix Weninger, Martin Wollmer, Jing Han, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Zhang, Zixing, Felix Weninger, Martin Wollmer, Jing Han, and Björn Schuller. 2017. "Towards intoxicated speech recognition." In *International Joint Conference on Neural Networks (IJCNN), 14-19 May 2017, Anchorage, AK, USA*, 1555–59. New York, NY: IEEE. https://doi.org/10.1109/ijcnn.2017.7966036.

Towards Intoxicated Speech Recognition

Zixing Zhang*, Felix Weninger†, Martin Wöllmer‡, Jing Han* and Björn Schuller*§
*Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany
Email: {zixing.zhang, jing.han}@uni-passau.de
†Nuance Communication, Ulm, Germany
Email: felix@weninger.de
‡BMW Group, Munich, Germany
Email: martin.woellmer@bmw.de
§Department of Computing, Imperial College London, London, UK
Email: schuller@ieee.org

***Abstract*—In a real-life scenario, the acoustic characteristics of speech often suffer from the variations induced by diverse environmental noises and different speakers. To overcome the speaker-related speech variation problem for Automatic Speech Recognition (ASR), many speaker adaptation techniques have been proposed and studied. Almost all of these studies, however, only considered the speakers' long-term traits, such as age, gender, and dialect. Speakers' short-term states, for example, affect and intoxication, are largely ignored. In this study, we address one particular speaker state, alcohol intoxication, which has rarely been studied in the context of ASR. To do this, empirical experiments are performed on a publicly available database used for the INTERSPEECH 2011 Speaker State Challenge, Intoxication Sub-Challenge. The experimental results show that the intoxicated state of the speaker indeed degrades the performance of ASR systems by a large margin for all of the three considered speech styles (spontaneous speech, tongue twisters, command & control). In addition, this paper further shows that multi-condition training can notably improve the acoustic model.**

I. Introduction

Improving the robustness of automatic speech recognition (ASR) against session variability caused by changes in the environment (e. g., different room acoustics) or the mental state of the speaker (e. g., angry speech) has been an essential research topic, and adaptation techniques as well as enhanced recognition architectures have been developed to cope with these influences. Apart from various types of background noise such as additive noise [1]–[5] or convolutional noise [6]–[9], adaptation to foreign and local regional accents has been considered: Non-native speakers often replace the unfamiliar phoneme in the target language which is missing in their native language phoneme dictionary [10], and accented speech is associated with a shift within the feature space [11]. Besides, physiological traits including age and gender seriously influence the performance of ASR [12], [13]. For instance, recognition of children's speech has been found to be highly challenging [14], [15]; furthermore, effects of gender are often mitigated by employing gender recognition prior to ASR [12], [16]. Apart from such long-term speaker traits, the significant influence of short-term speaker states on ASR accuracy has been demonstrated as well: Changing speech rates makes the mapping process between the acoustic signal and the phonetic categories more complex [17]. Besides, the speaker's emotional state is found to be significantly influential on the speech spectrum. For ASR, the recognition rate for the spontaneous emotionally coloured speech can be improved by using a language model based on increased representation of emotional utterances [18]. Similarly, a dynamic emotional adaptation has also been proposed for this issue in [19]. In addition to emotion, eating state has been investigated in [20] to show the performance impact on ASR systems. Finally, the impact of other 'intra-speaker' factors like speaker's health state, speaking style, social status, cultural background was covered in [21] for speaker-independent ASR.

However, another important intra-speaker factor, namely intoxication, has been largely neglected in the field of ASR as far as we know. With the expected increased usage of ASR in daily life situations, intoxication caused by alcohol or other drugs might become a common situation that ASR has to deal with. Even though the connection of acoustic parameters and intoxication has been analyzed, e. g., in [22], [23], there is little research on automatic recognition. Rather, there is a focus on forensic aspects, providing secondary evidence for alcohol impairment [24]. Works on automatic classification, in turn, mostly focus their attention on detecting intoxication, such as in the INTERSPEECH 2011 Speaker State Challenge [25]–[27]. In this paper, we investigate the challenge of alcohol intoxication to speaker-independent ASR. Two questions will be addressed in the paper: 1) Does alcohol intoxication state affect the performance of speaker-independent ASR? If yes, how serious is the influence? 2) If the answer to the first question is yes, can we enhance the robustness of the speech recognizer in this scenario? For such an investigation, we chose the most conventional ASR acoustic model (i. e., Gaussian Mixture Model and Hidden Markov Model [GMM/HMM]) and adaptation technique (i. e., Expectation-Maximization [EM]) as a start point of the experiments for the sake of reproducibility.

The remainder of this paper is structured as this: In Section II, the Alcohol Language Corpus (ALC) of genuine intoxicated speech is introduced. After that, the impact of alcohol intoxication on ASR is investigated in Section III-A by evaluating across different speech styles. Furthermore, acoustic model adaptation for alcohol intoxicated speech is discussed in Section III-B. Finally, the major findings are summarized

TABLE I
STATISTICS OF ALC TRAINING DATA: SOBER SUBSET (BAC PER MILL = 0), INTOXICATED SUBSET (BAC PER MILL $\neq$ 0), AND COMPLETE TRAINING SET (SOBER PLUS INTOXICATED). #AVG: AVERAGE WORD NUMBER PER UTTERANCE (UTT.)

Training Subsets	# utt.	# word	# Avg.
Sober	6 240	135 047	21.6
Intoxicated	3 120	63 980	20.5
Sober + Intoxicated	9 360	199 027	21.3

and future work is pointed out in Section IV.

II. THE ALCOHOL LANGUAGE CORPUS

The experiments described in the paper are based on a publicly available corpus – the Alcohol Language Corpus (ALC) containing 38 hours of genuine alcohol intoxicated and sober speech, which is distributed by the Bavarian Archive of Speech Signals (BAS) for unrestricted scientific and commercial usage [28]. This corpus has been used for the INTERSPEECH 2011 Speaker State Challenge (SSC) evaluating the automatic recognition of alcohol intoxication from speech [25].

For our experiments, as for the 2011 SSC, we use a gender balanced subset of the ALC with 154 speakers (77 male, 77 female). Speakers are within the age range of 21 to 75 years and were selected to ensure a balance of German dialects. All speakers are native German speakers. For our experiments, the recordings from 104 speakers are serving as training set (corresponding to the union of training and development set of the 2011 SSC), and the recordings from the other 50 speakers (2011 SSC test set) are serving as testing set, guaranteeing speaker independence and gender balance. Details of the data distribution can be found in Table I.

A controlled voluntary intoxication experiment was performed to create the ALC, supervised by the Munich Institute of Legal Medicine. The participants chose a blood alcohol concentration (BAC) that they wanted to attain in the experiment. To establish a solid ground truth for alcohol intoxication, a blood sample was taken 20 minutes after alcohol consumption. The speakers used for the 2011 SSC corpus, and hence our experiments, reached BACs ranging from 0.28 to 1.75 per mill (volume of alcohol per volume of blood, which is the legally binding unit of measurement in many countries). The intoxicated speech material in the ALC was obtained by a speech test which the speakers were asked to perform immediately after taking the blood sample. Since the speech test did not last longer than 15 minutes, it is ensured that the BAC throughout the speech test remains roughly equal to the measured BAC before the test. At least two weeks after the intoxicated speech test, each speaker returned to undergo a second recording in sober condition. The sober recordings were chosen to be roughly twice as long as the intoxicated recordings. Sober and intoxicated recordings were performed in the same acoustic environment and were conducted by the same BAS staff member to control for undesired influence factors on the acoustic features or dialogue behavior. The sampling rate of the recordings is 16 kHz.

Three different speech styles are included in the ALC: read speech, spontaneous speech, and command & control. The read speech comprises phrases often found in human-machine communication including connected digits and spelling, as well as tongue twisters which contain specific phonetic combinations that are expected to be hard to plan and produce especially under the influence of intoxication. Details can be found in [28]. Spontaneous speech consists of three monologues and four dialogues (twice as many in sober condition) with the recording supervisor, and is elicited by pictures to describe and personal questions, such as the description of the last vacation of the speaker, the most valued gift she or he had received, etc. Both monologues and dialogues have a length of at most 60 seconds each [28]. The command & control speech includes typical commands used in a driving environment, such as controlling of the air conditioning, street addresses for the GPS navigation, etc. There are both 'read' and 'spontaneous' commands; the former are taken from a real automobile prototype while the latter are elicited by asking the speaker to control the car with his/her own words in a specified driving situation [28]. All speakers are prompted with the same material.

III. RECOGNITION OF ALCOHOL INTOXICATED SPEECH

For the experiments, all features are extracted from frames of 25 ms length sampled at a rate of 10 ms. A Hamming window is applied to the frames before transformation to the spectral domain. From each frame, 12 cepstral mean normalized Mel-Frequency Cepstral Coefficient (MFCC) features together with energy as well as first and second order delta coefficients were extracted as feature vectors. Our ASR system is based on HMM using Baum-Welch reestimation for training and Viterbi decoding. 32 Gaussian mixture components are estimated for silence and 16 Gaussian mixture components for the other phonemes by iterative mixture splitting and re-training. Decision-tree clustered state-tied triphone models are created from 46 German monophone models including a model for hesitations.

As language model (LM), we employ a back-off bi-gram German language model trained on 170 million words of German newspaper texts (vocabulary size 151 k). We adapt the language model to the domain by adding all ALC training set utterances (199 k words) with double weight to the sentences used to train the LM. This small weight is chosen such as to include special vocabulary of the ALC without overfitting to particular patterns in the ALC speech tasks. The out-of-vocabulary rate of the ALC test set is at 1.25 %.

A. Impact of Intoxicated Speech

To evaluate the impact of alcohol intoxication on speech recognition, two testing scenarios are taken into account. In the first experiment, we define four testing subsets depending on various ranges of alcohol intoxication level: sober speech (BAC per mill = 0), 'mildly' intoxicated speech (BAC per mill $\in$]0, 0.5]), 'highly' intoxicated speech (BAC per mill $\in$]0.5,

TABLE II

STATISTICS OF THE WHOLE TESTING SET OF THE ALC AND THREE SUBSETS BY SPEECH STYLE (SP: SPONTANEOUS SPEECH; CC: COMMAND & CONTROL; TT: TONGUE TWISTER), WITH VARIOUS RANGES OF INTOXICATION LEVELS (SOBER, MILDLY INTOXICATED (ITX.), HIGHLY INTOXICATED, ALL OF INTOXICATED, AND ALL (SOBER & INTOXICATED)). #AVG.: AVERAGE WORD NUMBER PER UTTERANCE (UTT.).

Subsets	Total			SP			CC			TT		
	#utt.	#word	#Avg.	#utt.	#word	#Avg.	#utt.	#word	#Avg.	#utt.	#word	#Avg.
Sober	1 500	31 155	20.8	250	22 045	88.2	850	5 917	7.0	400	3 193	8.0
Mildly Itx.	120	2 659	22.2	20	1 891	94.6	68	509	7.5	32	259	8.1
Highly Itx.	1 380	27 048	19.6	230	18 036	78.4	782	5 921	7.6	368	3 085	8.4
Itx.	1 500	29 701	19.8	250	19 927	79.7	850	6 430	7.6	400	3 344	8.4
All	3 000	60 856	20.3	500	41 972	83.9	1 700	12 347	7.3	800	6 537	8.2

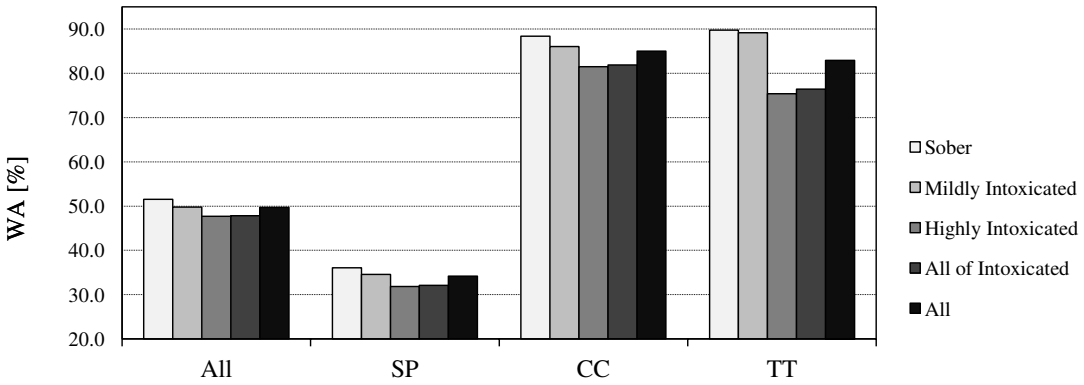

Fig. 1. Impact of alcohol intoxication on ASR performance for three speech styles. Word accuracy (WA, %) is given for the whole test set of the ALC (All) and three subsets corresponding to speech styles (SP: spontaneous speech; CC: command & control; TT: tongue twister). Different colours correspond to various ranges of intoxication level (see text for detailed explanation). Acoustic model training on sober data.

1.75]), and intoxicated speech in general (BAC per mill $\neq$ 0). The boundary between 'mildly' and 'highly' intoxicated is chosen by the common legal limit for driving. The left part of Table II displays the distribution of the test set with respect to the intoxication levels. The leftmost bar plot in Figure 1 depicts the performance on the testing subsets and the test set as a whole. It can be seen that the best performance is achieved by the subset produced by sober speech with 51.5 % word accuracy (WA), followed by mildly intoxicated speech with 49.8 % WA. The worst performance is observed by highly intoxicated speech with 47.9 % WA. From these results, we can conclude that the alcoholized speech significantly degrades the performance of speech recognition by 3.6 % WA absolute (one-side z-test, $p < .05$) if no further adaptation methods are implemented. This can be attributed to the effects of alcohol which leads to poor coordination and slurred speaking, etc.

In order to find which style of speech is affected most seriously, in a second experiment we subdivide the testing utterances into command & control speech, tongue twisters, and spontaneous speech. This subdivision is oriented on the expected difficulty of the speech planning, production and recognition tasks, and thus we subsume the command & control utterances and the numbers, address, and spelling utterances from the read speech part of the ALC, as these all correspond to possible applications in a car scenario, and are characterized by similar speech features: They usually consist of isolated words spoken with minimal pauses between them. In contrast, the 'tongue twisters' are complete sentences which are expected to produce continuous speech, whereby a number of difficult pronunciations like alveolar voiceless fricative alternating with the post-alveolar voiceless fricative are included [28]. Finally, the spontaneous speech as described in Section II, displays a variety of natural speech features such as fluent and disfluent speech, hesitations, sighs, laughter, repetitions, and so on. The right part of Table II shows the detailed distribution of the three speech styles. It can be seen that the command & control speech has the least average word number per each utterance, followed by the tongue twister speech. In contrast, for the spontaneous speech the average word number per utterance drastically rises up to 83.9.

Figure 1 shows the performance for each speech style. First, commands & control speech performs best with 85.0 % WA overall, followed by tongue twister speech with 82.9 % WA overall. However, the spontaneous speech performs worst, which is to be expected. Furthermore, we can see that obviously tongue twister speech is most seriously influenced by intoxication, as the WA drops from 89.7 % in sober condition (BAC per mill = 0) to 76.4 % in intoxicated condition (BAC per mill $\neq$ 0), which is a 13.3 % absolute decrease (one-side z-test, $p < .05$). Apparently, the alcohol severely impacts the

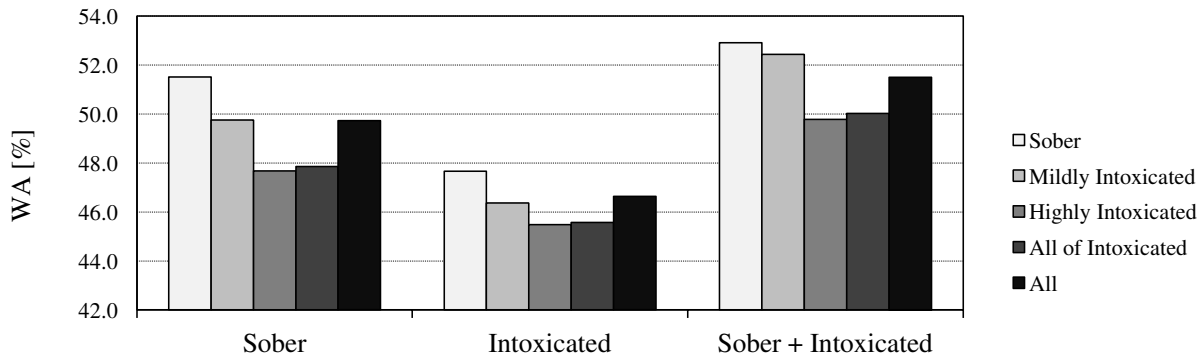

Fig. 2. Impact of training data for acoustic models. Word accuracy (WA, %) is given for subsets of the ALC test set corresponding to various intoxication levels (sober, mildly intoxicated, highly intoxicated, intoxicated), and the whole test set (All). Acoustic model training on the sober subset (BAC per mill = 0), intoxicated subset (BAC per mill $\neq$ 0), and joint sober and intoxicated subset.

speaker's speech planning. Compared to that, the performance on command & control speech decreases by 6.5 % WA absolute from the sober condition to the intoxicated condition, which is less degradation than for tongue twister speech. This can be explained by the fact that the command & control speech is often slowly pronounced, accentuated, or even hyper-articulated by intoxicated speakers. For spontaneous speech, the absolute drop in WA is only 4.0 %, yet, since the accuracy is generally lower, the relative decrease is similar to the other scenarios. Furthermore, it is evident that the state of mild intoxication just slightly affects the recognition performance compared to the sober condition.

B. Acoustic Model Training

From the first experiment, we conclude that there is a significant, sometimes even drastic, influence of alcohol intoxication on ASR, answering our first research question. Thus, we continue to investigate better model training for increasing the robustness for alcoholized speech. To obtain acoustic models for recognition of intoxicated speech, first, acoustic models trained on only alcohol intoxicated speech (BAC per mill $\in$]0, 1.75]) are evaluated. The results are shown in the middle bar plot of Figure 2. On average, the performance of a speech recognizer trained on intoxicated speech is worse than that of one trained on sober speech albeit this is partly due to less training data (cf. Table I). The performance for sober speech seriously drops to 47.7 % WA from 51.5 % WA, decreasing by 3.8 % absolute WA (one-side z-test, $p < .05$). In contrast, the performance for alcohol intoxicated speech is only lowered by 2.3 % absolute WA (one-side z-test, $p < .05$). This can be attributed to the matched speech condition. However, we expect that even with more training data, training with only intoxicated speech would overadapt to the intoxicated condition, at the expense of higher word error rate for sober speech.

Secondly, in order to improve the ASR performance, we integrate intoxicated speech into the baseline speech recognizer which is trained on sober data by performing additional EM iterations on intoxicated data, updating means and variances of Gaussian mixtures to capture possibly larger variation, as well as transition probabilities to model slurred speech. The larger size of joined sober and intoxicated speech data (cf. Table I) yields obviously higher word accuracies as shown in the right of Figure 2. Sober speech and intoxicated speech are now recognized with 52.9 % WA and 50.0 % WA compared to 51.5 % WA and 47.9 % WA in the baseline scenario, respectively, obtaining absolute increases of WA of 1.4 % and 2.1 %, respectively (one-side z-test, $p < .05$). The higher absolute improvement for intoxicated speech demonstrates that adding intoxicated speech can successfully improve the ASR system's performance for both sober speech and intoxicated speech.

IV. Conclusions

In this paper we investigated the impact of alcohol intoxication state on automatic speech recognition. The results show that, when faced with intoxicated speech, the performance of a speech recognizer trained on sober speech significantly degrades. The results also indicate that highly intoxicated speech impacts the performance more seriously than the mildly intoxicated speech which is recognized just slightly worse than the sober speech. Furthermore, we evaluated three styles of speech: spontaneous speech, command & control speech, and tongue twister speech. From the results, we found that the tongue twister speech is influenced most seriously, followed by command & control speech. This observation demonstrates that the alcohol intoxication affects the speaker's articulation, albeit accentuated and pronounced speech can alleviate this degradation. In turn, we found that training on intoxicated speech does not yield models that generalize well to sober speech.

To enhance the robustness of a speech recognizer for intoxicated speech, we added intoxicated speech to the baseline sober training set, yielding performance gains both for sober speech and alcohol intoxicated speech of 1.4 % and 2.1 % accuracy,

respectively. Thus, in contrast to training with intoxicated speech only, the increased performance for intoxicated speech is almost not at the expense of accuracy for sober speech.

For our future work, state-of-the-art speech recognition techniques will be investigated as well. For example, deep neural networks have been widely employed to distill the underlying representations of speech in an unsupervised manner and have frequently showed their effectiveness on speech recognition; memory-based neural networks have also been verified to be promising in capturing the long-term context information [29], [30], which is of significance for speech recognition. All these advanced deep learning techniques appear to be among the 'favorable weapons' in addressing the intoxicated speech recognition problem. In addition, other model adaptation techniques, such as Maximum Likelihood Linear Regression, are worth being evaluated as well.

ACKNOWLEDGMENT

This work was supported by the European Union's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and Horizon 2020 Programme through the Research Innovation Action No. 645378 (ARIA-VALUSPA).

REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr 2014.

[2] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 3593–3597.

[3] H. B. D. Sorensen, "Noise-robust speech recognition using a cepstral noise reduction neural network architecture," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, Seattle, WA, 1991, pp. 795–800.

[4] J. Gao, J. Du, C. Kong, H. Lu, E. Chen, and C. H. Lee, "An experimental study on joint modeling of mixed-bandwidth data via deep neural networks for robust speech recognition," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, 2016, pp. 588–594.

[5] S. Squartini, E. Principi, R. Rotili, and F. Piazza, "Environmental robust speech and speaker recognition through multi-channel histogram equalization," *Neurocomputing*, vol. 78, no. 1, pp. 111–120, Feb 2012.

[6] Z. Zhang, J. Pinto, C. Plahl, B. Schuller, and D. Willett, "Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 525–533, Aug 2014.

[7] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, Apr 2015.

[8] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNS for distant speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5755–5759.

[9] R. Rotili, E. Principi, S. Squartini, and B. Schuller, "A real-time speech enhancement framework in noisy and reverberated acoustic scenarios," *Cognitive Computation*, vol. 5, no. 4, pp. 504–516, Dec 2013.

[10] M. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 4398–4401.

[11] D. V. Compernolle, J. Smolders, P. Jaspers, and T. Hellemans, "Speaker Clustering for Dialectic Robustness in Speaker Independent Recognition," in *Proc. EUROSPEECH*, Genova, Italy, 1991, pp. 723–726.

[12] T. Zoughi and M. M. Homayounpour, "Gender aware deep boltzmann machines for phone recognition," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1–5.

[13] R. Vipperla, S. Renals, and J. Frankel, "Ageing voices: The effect of changes in voice parameters on asr performance," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 1–11, Feb 2010.

[14] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, Nov 2003.

[15] R. Sinha, S. Shahnawazuddin, and P. S. Karthik, "Exploring the role of pitch-adaptive cepstral features in context of children's mismatched ASR," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2016, pp. 1–5.

[16] D. Deiv, Gaurav, and M. Bhattacharya, "Automatic gender identification for hindi speech recognition," *International Journal of Computer Applications*, vol. 31, no. 5, pp. 1–8, Oct 2011.

[17] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, 2004, pp. 549–552.

[18] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: clarifying the issues and enhancing performance," *Neural Networks*, vol. 18, no. 4, pp. 437–444, May 2005.

[19] S. Steidl, A. Batliner, D. Seppi, and B. Schuller, "On the impact of children's emotional speech on acoustic and language models," *EURASIP Journal Audio Speech Music Processing*, vol. 2010, pp. 1–14, Jan 2010.

[20] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. E.-D. Mousa, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on ASR performance," *PloS one*, vol. 11, no. 5, pp. 1–24, May 2016.

[21] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, Nov 2007.

[22] D. B. Pisoni and P. C. Martin, "Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses," *Alcoholism: Clinical and Experimental Research*, vol. 13, no. 4, pp. 577–587, Aug 1989.

[23] H. Hollien, G. Dejong, C. A. Martin, R. Schwartz, and K. Liljegren, "Effects of ethanol intoxication on speech suprasegmentals," *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3198–3206, Dec 2001.

[24] M. Brenner and J. Cash, "Speech analysis as an index of alcohol intoxication - the Exxon Valdez accident," *Aviation, Space, and Environmental Medicine*, vol. 62, pp. 893–898, Sep 1991.

[25] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. INTERSPEECH 2011*, Florence, Italy, 2011, pp. 3201–3204.

[26] F. Weninger and B. Schuller, "Fusing utterance-level classifiers for robust intoxication recognition from speech," in *Proc. International Conference on Multimodal Interaction (ICMI)*, Alicante, Spain, 2011, pp. 1–2.

[27] Z. Zhang, F. Weninger, and B. Schuller, "Towards automatic intoxication detection from speech in real-life acoustic environments," in *Proc. ITG Symposium on Speech Communication*, Braunschweig, Germany, 2012, pp. 1–4.

[28] F. Schiel and C. Heinrich, "Laying the Foundation for In-Car Alcohol Detection by Speech," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 983–986.

[29] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, Dec 2016, in press.

[30] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, 5 pages.