EMOTION BIAS IN AUTOMATIC SPEECH RECOGNITION

Lara-Sophie Christmann[1]

[1]*TU Berlin*
christmann@campus.tu-berlin.de

Kurzfassung: In this paper, we investigate the effect of emotions on an established automatic speech recognition system using five emotional speech databases covering English, German, and Italian language. We computed the word error rates and the significance of the ratio between the correctly and incorrectly recognized words per emotion category. Results showed a strong bias with an increase in word error rates of up to +73.7% when compared to neutral speech. The correlation between emotional categories and error rates was significant at $p = 0.001$ for all datasets. We further tested the applicability of an existing CycleGAN for emotional speech conversion as a preprocessing step to transform speech to neutral state. The de-emotionalized speech produced by the trained networks was retested for recognition rates in comparison to emotionalized neutral speech.

1 Introduction

Automatic speech recognition systems (ASRSs) have been significantly improved with the development of machine learning and Artificial Intelligence. Nonetheless, a prominent issue that ASRSs must contend with is the sheer amount of data necessary to achieve valuable results. The scientific requirements for such datasets are hard to fulfill. A valid set should cover all different data variations in a balanced and sufficient number. When those requirements are not met, the model is biased. A known example in speech recognition is gender bias in earlier versions of Amazon's Alexa ASRS as well as YouTube's automatic speech-to-text system [1, 2]. Another easily overlooked issue is real-life speech recognition scenarios, where ASRSs encounter emotional speech. Exemplary situations of this are stressed drivers talking to their navigational systems, elderly interacting with smart care robots or emotionally moved users talking to their smartphones or home assistants. Since spoken emotions distort speech from a neutral state, they can interfere with detection quality. While ASRSs are exposed to these signals, little research exist about how well they are adapted to them. Enough emotional samples would have to exist to train larger ASRSs to build a well-balanced system capable of handling speech recognition across all emotional vocal states equally well. If not, the detection quality decreases. As a result, the user experience can suffer if the ASRS performs poorly during emotional vocalization especially in stressful situations. Some users might adapt to the bias by mimicking more neutral speech, others might lose interest in interacting with such systems. Some research has been directed towards building emotion speech recognition systems based on emotion-oriented language models and dictionaries [3, 4] or by combining emotion recognition and emotion dependent models [5]. Others try to abstract the speech signal in the source domain while preserving the verbal content to avoid emotion bias [6], or construct a recognition system which combines ASRS and emotion masking [7].

2 Bias in Automatic Speech Recognition

The existence of bias is a recurring topic in ASRS research. One of the first discovered biases in ASRS was a gender bias against female speakers. In 2017, Tatman and Kasten [8] conducted a study on Youtube's automatic captions for female and male speakers using five different English dialects. The data was from the accent tag challenge, where speakers annotated their language background on their videos. By filtering Youtube content for these tagged videos the researchers created their own dataset resulting in eight speakers from California, Georgia, New England, New Zealand, and Scotland. The suggested captions from Youtube were then checked by a phonetically trained expert for each dialect. The study evaluation showed that the word error rate (WER) differences across all English dialects were biased against women. Furthermore, a regional bias against the Scottish speakers yielded a significant WER increase. The researchers additionally showed a correlation between dialect and gender, where discrepancies were worse for New Zealand and New-England based speakers.

In 2020, Koenecke et al. proved a racial bias against black speakers. In their paper [9] they tested white and black speaker recordings on Amazon, Apple, Google, IBM and Microsoft's ASRS and found substantial proof of bias. The data basis were two new corpora of conversational recordings, the Corpus of Regional African American Language (CORAAL) [10] and the Voices of California (VOC) [11] database. The CORAAL contains sociolinguistic interviews with a multitude of speakers from black communities speaking African American vernacular English (AAVE), which is a term describing the ethnolect spoken by African American urban communities in the United States and parts of Canada. The VOC contains interviews conducted with people across the state of California within a dyadic setting of interviewer and interviewee. For the bias study a subset of the VOC containing only the interviewee snippets was compiled. Finally, the study compared the WERs for white and black speakers across all five tested ASRSs and the average WER per category. The results clearly showed a rise in detection errors for black speakers with rates being approximately twice as high than for white speakers. The aggregated WER across the ASRS for black speakers was 0.35 while it was only 0.19 for white speakers.

Kostoulas et al. evaluted the performance of the open-source smart-home ASRS dialogue system Sphinx III on emotional speech. They used an acoustic model based on the Wall Street Journal database and compared WERs of the Speech and Transcripts database. The database contains fourteen emotional categories and 'Neutral'. The researchers showed that the 'Neutral' category yielded the lowest WER and suggested using an emotion adaptive ASRS to increase robustness [12].

3 Automatic Speech Recognition Evaluation of Emotional Speech

Even though established ASRSs have been known to be biased few research has been conducted on the performance of those systems on emotional data. Most research has been concerned with constructing own systems instead of evaluating established models. In the following, we explore the affect of emotional speech on the Google Cloud Speech-to-Text API [13].

3.1 Experiment Setup

To test whether emotions affect the detection quality we constructed the following test scenario: We chose five different emotional speech datasets, EmoDB [14], RAVDESS [15], CREMA-D [16], EmoV-DB [17] and Emovo [18] covering German, English and Italian language and evaluated each emotion category within each set for its performance. All datasets contain acted

speech recorded in a studio. Dimensional factors such as intensity were not considered. We decided to thoroughly test the ASRS on a multitude of databases to exclude verbal content dependency of the results. We split the data by emotion category and created lookup files with lists of file identification strings and the recorded sentence. Next, each speech file from an emotion was processed by the Google service. Except for the language code of the dataset, no further adjustments were made to the API requests. All returned strings of a category were stored. In parallel, we build a reference string using the file identifiers to extract the original content of the speech from the lookup files. If an unrecognized error occurred, the sentence was flagged but not included. The Levenshtein distance was used to compute the number of insertions, deletions and replacements and the WER based on these strings. When applied at the word level, it counts the errors per category by finding the minimum path between the sentences [19]. Next, we constructed contingency tables for each dataset over all counted words by the algorithm for the discrete variables emotion categories and word categories. Word category has two possible values: 'error' or 'correct'. Since deletion errors and partially insertion errors yielded low absolute sums all three errors were pooled into the 'error' word category. Then, χ^2 was computed for all contingency tables to perform a Pearson's χ^2 test to examine the independence for the discrete distributions of word categories and emotion. We formulated the nullhypothesis: Emotion category and word category ('error' and 'correct') are independent variables. The alternate hypothesis was: Emotion and word category are dependent on each other.

3.2 Results

The WERs of emotion categories within datasets (compare Tables 1 to 8) show that emotion categories performed siginificantly worse than 'Neutral'. This finding was confirmed by the χ^2 test which proved dependency between the word categories and emotions. The null hypothesis can be rejected at a significance level of $p = 0.001$, since, for all databases under consideration, $p < 0.0001$ (compare Table 9). Since for some emotions noticeably low WERs occurred, partitioned χ^2 was performed in a next step. This should test whether significance holds for single emotion categories with a WER difference below 5% in comparison to 'Neutral'. At first, EmoDB's categories 'Disgust' and 'Sad' were under observation. Each one was tested solitarily against 'Neutral'. Here, the results did not indicate significance. Secondly, for CREMA-D, 'Anger' was tested against 'Neutral' and confirmed significance at $p = 0.001$. Thirdly, for Emovo 'Disgust' did not prove significance against 'Neutral'. Lastly, for EmoV-DB, all categories affirm strong significance even for WERs below 5% difference.

Tabelle 1 – Distribution of errors for EmoDB.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Happy	629	45	17	2	**9.5 %**	+5.5 %
Fear	550	78	42	0	**19.1 %**	+15.1 %
Disgust	407	23	2	2	**6.2 %**	+2.2 %
Sad	600	21	10	0	**5 %**	+1 %
Anger	1107	96	39	3	**11.4 %**	+7.4 %
Bore	711	57	31	1	**11.6 %**	+ 7.6 %
Neutral	731	25	5	0	**4 %**	-

In conclusion, significance was confirmed for all contingency tables at set level. For smaller WER differences, significance could not be found within the German dataset EmoDB for 'Disgust' and 'Sad' when tested singularly against 'Neutral'. For Emovo significance could not

Tabelle 2 – Distribution of errors for RAVDESS. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Happy	942	102	102	1	**19.6 %**	+16.6 %
Fear	848	122	182	55 (54)	**35 %** (31.4 %)	+32.0 %
Disgust	933	92	121	14 (12)	**21.9 %** (20.9 %)	18.9 %
Sad	920	117	115	13 (1)	**23.3 %** (22.4 %)	+20.3 %
Anger	1039	50	63	1	**10.50 %**	+7.5 %
Calm	1046	67	39	0	**9.5 %**	+6.5 %
Surprise	941	98	113	1	**20.4 %**	+17.4 %
Neutral	559	12	5	0	**3.0 %**	-

Tabelle 3 – Distribution of errors for CREMA-D. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Happy	4265	816	1542	536 (11)	**51.5 %** (46.5 %)	+34.7 %
Fear	3566	926	2131	964 (9)	**73.7 %** (68.1 %)	+56.9 %
Disgust	3652	852	2119	1020 (8)	**72.2 %** (66 %)	+55.4 %
Sad	3530	845	2248	1204 (5)	**77 %** (70.7 %)	+60.2 %
Anger	5387	654	582	74 (14)	**21.4 %** (20.6 %)	+4.6 %
Neutral	5244	359	463	145 (3)	**16.8 %** (14.7 %)	-

be shown for the category 'Disgust' against 'Neutral'.

4 Experimental Bias Removal via a Generative Adversarial Network

Emotion in speech can also be seen as a distortion that functions as noise. Improving noise conditions for speech detection is an ongoing and important research task. However, in the case of emotions, the noise is embedded within the voice itself and therefore can not be removed easily with classic methods of signal processing. State-of-the-art Artificial Intelligence offers new approaches to de-emotionalize speech.

In the past years, GANs have been a major research trend. Common areas of interest are image or video manipulation such as style transfer [20] or image denoising [21]. Dumpala et al. used GANs for voice conversion. The authors found that perturbations in speech, either as background noise or introduced by the speaker such as in laughing, degraded ASRS performance. Therefore, they suggested a CycleGAN to reduce perturbations in speech and improve ASRS performance [22].

Tabelle 4 – Distribution of errors for Emovo. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Joy	458	58	36	12 (2)	**20.1 %** (18.50 %)	+9.2 %
Fear	452	64	36	3	**19.80 %**	+8.9 %
Disgust	496	41	15	5	**11.30 %**	+0.4 %
Sad	455	58	39	24 (9)	**20.3 %**	+9.4 %
Surprise	465	49	38	12 (1)	**18.8 %** (17.1 %)	+7.9 %
Anger	462	59	31	0	**17.3 %**	+6.4 %
Neutral	495	42	15	2	**10.9 %**	-

Tabelle 5 – Distribution of errors for EmoV-DB for speaker Jenie. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Amused	1355	391	232	28 (24)	**36.7 %** (36.6 %)	+18.8 %
Angry	3295	752	328	59	**27.7 %**	+9.8 %
Disgusted	1368	267	75	19	**21.8 %**	+3.9 %
Sleepy	3070	812	235	60	**28.1 %**	+10.2 %
Neutral	3094	480	110	60 (49)	**17.9 %** (17.6 %)	-

Tabelle 6 – Distribution of errors for EmoV-DB for speaker Bea. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Amused	1676	651	338	39	**43.4 %**	+19.4 %
Angry	1852	651	238	50	**36.8 %**	+12.8 %
Disgusted	2221	545	207	47	**28.4 %**	+4.4 %
Sleepy	2973	1015	408	71 (60)	**36.8 %** (36.6 %)	+12.8 %
Neutral	2495	551	136	58	**24 %**	-

4.1 Emotional Voice Conversion

The CycleGAN architecture has its advantages in domains with sparse or unpaired datasets. In 2020, Zhou et al. presented a CycleGAN for spectrum and prosody conversion of emotional speech. Their architecture has two separate sub-CycleGANs, one for converting mel-cepstral coefficients (MCEPs) for spectrum conversion and the other for prosody conversion of the fundamental frequency F0. The prosody CycleGAN is an extension of the basic architecture, which means that emotional voice conversion can be achieved by only training the spectrum conversion [23]. The basis for this model was first presented by Lei Mao on Github and later adopted by Gao et al. [24]. Each speech signal is decomposed into F0, its timeaxis, spectral envelope and aperiodicity using WORLD [25]. Then, the spectral envelope is encoded using pyworld into a 24-dimensional MCEPs vector. For F0, mean and standard deviation are computed across all training samples. The spectrum CycleGAN is trained on the 24-dimensional MCEPs feature vector. The fundamental frequency of the test signal's pitch is adjusted to the target data by applying logarithm Gaussian normalization. Finally, the time-domain speech signal is reconstructed using the aperiodicity of the input signal, the transformed MCEPs feature vector and the adjusted F0. The WORLD vocoder is capable of resynthesizing speech using these features.

In an exploratory experiment, we tested the CycleGAN's ability to transfer emotional speech to neutral state to test the effects of the de-emotionalized speech on the Google Cloud Speech-to-Text API. We trained a spectrum CycleGAN on EmoV-DB's speaker Jenie for each emotion to 'Neutral'. The data was split in an 80 to 20 ratio into training and testing sets and the hyperparameters remained as suggested by the authors.

Tabelle 7 – Distribution of errors for EmoV-DB for speaker Josh. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Amused	744	859	1087	145 (10)	**119.6 %** (121.3 %)	+74.7 %
Sleepy	853	787	727	54 (15)	**92.6 %** (92.4 %)	+47.7 %
Neutral	1642	801	271	47 (40)	**44.9 %** (44.8 %)	-

Tabelle 8 – Distribution of errors for EmoV-DB for speaker Sam. Numbers in brackets do not consider completely unrecognized sentences.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Amused	1520	1441	1478	125 (26)	**98.6 %**	+58.5 %
Angry	2404	1293	449	57	**47.9 %**	+7.8 %
Disgust	2342	1270	797	49 (44)	**57.8 %** (57.7 %)	+17.7 %
Sleepy	1625	1519	1243	41 (18)	**88 %** (87.9 %)	+47.9 %
Neutral	2789	1216	353	61	**40.1 %**	-

Tabelle 9 – Significance statistics of word category ('error' or 'correct') and emotion.

Database	**Degree of Freedom**	χ^2	p-**Value**
EmoDB	6	261.508	<0.0001
RAVDESS	7	330.815	<0.0001
CREMA-D	5	4420.1	<0.0001
Emovo	6	39.3065	<0.0001
EmoV-DB Jenie	4	197.197	<0.0001
EmoV-DB Bea	4	210.485	<0.0001
EmoV-DB Josh	2	679.851	<0.0001
EmoV-DB Sam	4	1203.09	<0.0001

4.2 Performance Evaluation

Due to the CycleGAN architecture, we obtained transformations in both directions, namely neutralized emotional speech and emotionalized neutral speech. We used only transformed speech for the evaluation for better comparability and fairness regarding the quality of the produced speech. Both groups of speech were tested on the Google Cloud Speech-to-Text API. The WERs for all transformed signals remained higher than for the original dataset, which was not surprising, since we had little training data and expected some additional quality loss. Still, for the two categories 'Angry' and 'Disgusted', which show low to no additional perturbations such as yawning or laughter (see Table 10), the significance was inverted. The de-emotionalized speech yielded noticeably lower WERs than their emotionalized counterparts.

Tabelle 10 – Distribution of errors for EmoV-DB's speaker Jenie transformed by the CycleGAN.

	Correct	**Substitution**	**Insertion**	**Deletion**	**WER**	**Diff**
Ang to Neu	590	187	69	9	33.7 %	
Neu to Ang	411	195	130	9	54.3 %	+20.6 %
Dis to Neu	269	57	17	4	23.6 %	
Neu to Dis	520	154	62	10	33 %	+ 9.4 %
Sle to Neu	518	202	111	14	44.6 %	+7 %
Neu to Sle	493	179	64	15	37.6 %	
Amu to Neu	255	78	68	2	44.2 %	+2.5 %
Neu to Amu	468	188	80	9	41.7 %	

5 Conclusion

The WER distributions of the five databases under consideration show that the Google Cloud Speech-to-Text API performs worse for all emotional categories in comparison to neutral speech.

The Pearson's χ^2 test strongly supports the evidence that emotions and detection rates are dependent across all datasets with a few database dependent category exceptions. Thus, emotions do affect ASRS performance to a significant degree here, which confirms an emotion bias. Further research must be conducted to see if these results generalize to other large ASRSs and less controlled experiment setups, such as unacted emotional speech.

We further explored an emotion transfer CycleGAN to remove the bias before forwarding the speech to the Google Cloud Speech-to-Text API. The results indicated potential for the CycleGAN as a preprocessing step to improve robustness, but further experiments are needed. Finding a fitting database was a major issue for this task, which had a strong impact on the results. WERs of the transformed speech remained higher per emotion category than the original data.

Literatur

[1] DIZON, G.: *Evaluating intelligent personal assistants for l2 listening and speaking development. Language Learning & Technology*, 24(1), 2020.

[2] TATMAN, R.: *Gender and dialect bias in youtube's automatic captions*. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, S. 53–59. 2017.

[3] ATHANASELIS, T., S. BAKAMIDIS, I. DOLOGLOU, R. COWIE, E. DOUGLAS-COWIE, und C. COX: *Asr for emotional speech: clarifying the issues and enhancing performance*. *Neural Networks*, 18(4), S. 437–444, 2005.

[4] ATHANASELIS, T., S. BAKAMIDIS, und I. DOLOGLOU: *Recognising verbal content of emotionally coloured speech*. In *2006 14th European Signal Processing Conference*, S. 1–4. IEEE, 2006.

[5] SCHULLER, B., J. STADERMANN, und G. RIGOLL: *Affect-robust speech recognition by dynamic emotional adaptation*. In *Proc. Speech Prosody 2006, Dresden*. 2006.

[6] MUKAIHARA, K., S. SAKTI, und S. NAKAMURA: *Recognizing emotionally coloured dialogue speech using speaker-adapted dnn-cnn bottleneck features*. In *International Conference on Speech and Computer*, S. 632–641. Springer, 2017.

[7] BASHIRPOUR, M. und M. GERAVANCHIZADEH: *Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments*. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1), 2018.

[8] TATMAN, R. und C. KASTEN: *Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions.* In *INTERSPEECH*, S. 934–938. 2017.

[9] KOENECKE, A., A. NAM, E. LAKE, J. NUDELL, M. QUARTEY, Z. MENGESHA, C. TOUPS, J. R. RICKFORD, D. JURAFSKY, und S. GOEL: *Racial disparities in automated speech recognition*. *Proceedings of the National Academy of Sciences*, 117(14), S. 7684–7689, 2020.

[10] KENDALL, T. und C. FARRINGTON: *The corpus of regional african american language*. *Version*, 6, 2018.

[11] LINGUISTICS, S.: *Voices of california*. 2020. URL `http://web.stanford.edu/dept/linguistics/VoCal/`. [Accessed Jan. 21, 2021].

[12] KOSTOULAS, T., I. MPORAS, T. GANCHEV, und N. FAKOTAKIS: *The effect of emotional speech on a smart-home application*. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, S. 305–310. Springer, 2008.

[13] GOOGLE: *Google cloud speech-to-text api*. 2021. URL `https://cloud.google.com/speech-to-text`. [Accessed Jan. 21, 2021].

[14] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. F. SENDLMEIER, und B. WEISS: *A database of german emotional speech*. In *Ninth European Conference on Speech Communication and Technology*. 2005.

[15] LIVINGSTONE, S. R. und F. A. RUSSO: *The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english*. *PloS one*, 13(5), 2018.

[16] CAO, H., D. G. COOPER, M. K. KEUTMANN, R. C. GUR, A. NENKOVA, und R. VERMA: *Crema-d: Crowd-sourced emotional multimodal actors dataset*. *IEEE transactions on affective computing*, 5(4), S. 377–390, 2014.

[17] ADIGWE, A., N. TITS, K. E. HADDAD, S. OSTADABBAS, und T. DUTOIT: *The emotional voices database: Towards controlling the emotion dimension in voice generation systems*. *arXiv preprint arXiv:1806.09514*, 2018.

[18] COSTANTINI, G., I. IADEROLA, A. PAOLONI, und M. TODISCO: *Emovo corpus: an italian emotional speech database*. In *International Conference on Language Resources and Evaluation (LREC 2014)*, S. 3501–3504. European Language Resources Association (ELRA), 2014.

[19] LEVENSHTEIN, V. I.: *Binary codes capable of correcting deletions, insertions, and reversals*. In *Soviet physics doklady*, Bd. 10, S. 707–710. 1966.

[20] ZHANG, L., Y. JI, X. LIN, und C. LIU: *Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan*. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, S. 506–511. 2017. doi:10.1109/ACPR.2017.61.

[21] CHEN, J., J. CHEN, H. CHAO, und M. YANG: *Image blind denoising with generative adversarial network based noise modeling*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, S. 3155–3164. 2018.

[22] DUMPALA, S. H., I. SHEIKH, R. CHAKRABORTY, und S. K. KOPPARAPU: *Cycle-consistent gan front-end to improve asr robustness to perturbed speech*. In *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language*. 2018.

[23] ZHOU, K., B. SISMAN, und H. LI: *Transforming spectrum and prosody for emotional voice conversion with non-parallel training data*. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, S. 230–237. 2020.

[24] GAO, J., D. CHAKRABORTY, H. TEMBINE, und O. OLALEYE: *Nonparallel emotional speech conversion*. 2019.

[25] MORISE, M., F. YOKOMORI, und K. OZAWA: *World: a vocoder-based high-quality speech synthesis system for real-time applications*. *IEICE TRANSACTIONS on Information and Systems*, 99(7), S. 1877–1884, 2016.