# Emotion Impacts Speech Recognition Performance

**Rushab Munot,**
University of Pennsylvania
rushab@cis.upenn.edu

**Ani Nenkova,**
University of Pennsylvania
nenkova@cis.upenn.edu

## Abstract

It has been established that the performance of speech recognition systems depends on multiple factors including the lexical content, speaker identity and dialect. Here we use three English datasets of acted emotion to demonstrate that emotional content also impacts the performance of commercial systems. On two of the corpora, emotion is a bigger contributor to recognition errors than speaker identity and on two, neutral speech is recognized considerably better than emotional speech. We further evaluate the commercial systems on spontaneous interactions that contain portions of emotional speech. We propose and validate on the acted datasets, a method that allows us to evaluate the overall impact of emotion on recognition even when manual transcripts are not available. Using this method, we show that emotion in natural spontaneous dialogue is a less prominent but still significant factor in recognition accuracy.

## 1 Introduction

Alexa and Google Home are becoming increasingly popular, their use spanning a range of applications from reducing loneliness in the elderly (Reis et al., 2017; Ferland et al., 2018) to child entertainment and education (Druga et al., 2017). As these conversational agents become commonplace, people are likely to express emotion during their interactions, either because of their perception of the agent or because of the emotion-eliciting situations in which the agent is deployed.

In this paper, we set out to study the extent to which emotional content in speech impacts speech recognition performance *of commercial systems*. Similar studies have been conducted in the past to study how recognition varies with gender and dialect (Adda-Decker and Lamel, 2005; Tatman, 2017), lexical content (Goldwater et al., 2010), topical domain (Traum et al., 2015) and delivery style (Siegler and Stern, 1995; Nakamura et al., 2008). A number of studies have studied the impact of stress (Hansen and Patil, 2007; Bou-Ghazale and Hansen, 2000; Steeneken and Hansen, 1999; Hansen, 1996) and emotional factors (Polzin and Waibel, 1998; Kostoulas et al., 2008; Benzeghiba et al., 2007) on speech recognition. Multiple studies (Byrne et al., 2004; Athanaselis et al., 2005; Pan et al., 2006; Meng et al., 2007; Ijima et al., 2009; Sun et al., 2009; Sheikhan et al., 2012) tried to improve upon speech recognition accuracies for emotional speech. However, these studies were carried out with older recognition systems. Recently automatic speech recognition has seen unprecedented gains in accuracy. Yet our work shows that emotional content still poses problems to speech recognition systems.

In our work we seek to quantify the influence of emotion on recognition accuracy for three commercial systems, on several datasets. We start out with two datasets of acted emotion, which are in some respects ideal for the task because the spoken content is constrained to pre-selected utterances and thus manual transcription is not required. In addition, the lexical content for each emotion is identical, so no special adjustment for that confounding factor is needed in the acted corpora.

At the same time, it is important to validate these results on spontaneous, more natural exchanges, so we also present results on such a corpus of emotion in spontaneous speech. As the speech becomes more natural, it becomes harder to obtain large manual transcripts for a large portion of the data to carry out the studies that we present, so we also validate an alternative method for finding factors that influence the performance of commercial systems, relying on agreement between systems rather than manual transcripts. We present convincing evidence that the approach is a



*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 16–21
Minneapolis, Minnesota, June 3 - 5, 2019. ©2017 Association for Computational Linguistics

reasonable approximation and it can be used for broader studies on factors influencing automatic speech recognition. Here, we apply the method to analyze data from a spontaneous emotional speech corpus.

## 2 Related Work

There has been much research in the field of emotion recognition from speech. However, relatively less research has been conducted on how emotion affects Automatic Speech Recognition (ASR).

(Polzin and Waibel, 1998) study the variation in word error rate with different emotions - NEUTRAL, ANGER, HAPINESS, AFRAID and SADNESS. They observed that SADNESS and AFRAID/FEAR perform worst while NEUTRAL and ANGER perform best. They integrate prosodic features into the model using Hidden Markov Models to first disambiguate the emotional state of the speaker, and then use emotion specific ASR models for transcription. They report a significant increase in ASR performance.

(Kostoulas et al., 2008) conduct a similar study over a much wider range of emotions (About 15 emotions) on the Wall Street Journal database with Sphinx III as the ASR system and report a large variance in the WER across emotions, ranging from about 6% for NEUTRAL to about 44% for HOT ANGER.

(Athanaselis et al., 2005) extract an emotionally *colored* subset of the British National Corpus (BNC) and append it multiple times to the BNC before training an emotionally-enhanced ASR system.

(Sheikhan et al., 2012) propose that the emotion in speech leads to changes in Mel-frequency cepstral coefficients (MFCC) and thus propose neutralizing MFCCs by *warping* the first three formant frequencies and conduct their experiments to analyze improvement in ASR performance for the emotions ANGER and HAPPINESS.

In this work, we analyze the performance of multiple modern commercial ASR systems on emotional speech. We further quantify the correlation between the Word Error Rate and emotion. We also compare the dependence of ASR performance on other factors - speaker identity and spoken content with the dependence on emotion.

## 3 Datasets and APIs

We use three acted Emotion datasets: CREMA-D (Cao et al., 2014), RAVDESS (Steven R. Livingstone1, 2018) and MSP-IMPROV (Busso et al., 2017). CREMA-D has 12 sentences recorded by 91 actors in 6 different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad), for a total of 7,442 utterances in the dataset.[1]

RAVDESS has just two sentences, which are very similar to each other (*Kids are talking by the door* and *Dogs are sitting by the door*), recorded by 24 actors in 8 emotions with the addition of Surprised and Calm. RAVDESS has a total of 1,440 utterances, and each sentence is recorded in two intensities with two repetitions of each.

In addition to actors recording sentences in a pre-specified emotion, the MSP-IMPROV dataset contains 'improvised recordings', where actors converse to induce the desired emotion. In these interactions, there is at least one emotional rendition of the target utterance but other utterances may be emotionally neutral. MSP-IMPROV is comprised of 1,272 utterances distributed over 20 target sentences and four emotions (Neutral, Anger, Happy, Fear). We refer to this part of the corpus as *MSP-IMPROV Target*, where we only concern ourselves with recognizing the target sentence. We refer to the set of complete conversations as the *MSP-IMPROV Dialogue* Corpus. Manual transcripts are not available for this part. MSP-IMPROV has 1,085 complete conversations.

Commercial systems used for speech recognition are IBM Watson Speech-to-Text, Google Cloud Speech-to-Text and Amazon Transcribe. [2]. We will denote the APIs simply by IBM, GCP and AWS respectively.

## 4 Evaluation Metrics

We report two measures of automatic speech recognition performance: the Word Error Rate (WER) and the percentage of completely recognized sentences (CR). Minor semantic and grammatical errors are ignored by manually listing semantically equivalent sentences for computing CR. We report $1 - CR$ instead of $CR$ to maintain consistency with $WER$ interpretation, lower

[1] One of the sentences is recorded in three different intensities.

[2] **Websites:** https://www.ibm.com/watson/services/speech-to-text/, https://aws.amazon.com/transcribe/ and https://cloud.google.com/speech-to-text/ respectively

the better. We then perform an ANOVA analysis to determine the statistical significance of each factor in determining the WER.

The dialogues in the MSP-IMPROV corpus do not contain manual transcripts. We propose a metric to analyze the relative performance of different systems with varying emotions when manual transcripts are not available.

We calculate the performance of every system relative to other systems and then report the average of these cross-comparisons. This method accurately predicts the relative performance on emotional and neutral speech consistent with WER/CR results on the other corpora for which transcripts are available.

| Dataset | Metric | IBM | AWS | GCP |
|---|---|---|---|---|
| CREMA-D | WER | **10.00** | 13.09 | 18.80 |
| | 1-CR | **24.72** | 39.20 | 41.78 |
| RAVDESS | WER | **5.08** | 13.31 | 6.19 |
| | 1-CR | **9.17** | 56.38 | 15.49 |
| MSP-IMPROV Target | WER | 13.90 | **9.21** | 12.56 |
| | 1-CR | 38.76 | **35.72** | 39.54 |

Table 1: Overall Performance of IBM, AWS and GCP

## 5 Observations

### 5.1 Variations across various factors

The overall performance of the APIs on the three datasets is given in Table 1. IBM performs best on 2 out of 3 datasets (CREMA-D and RAVDESS). AWS is, however, more consistent across datasets.

Figure 1 shows how WER varies with emotion. On CREMA-D, NEUTRAL speech is recognized more accurately than emotional speech, with ANGER most accurately recognized among the emotions, while SADNESS and FEAR are poorly recognized. Similarly on RAVDESS, FEAR has the worst WER. NEUTRAL and ANGER are recognized better than other emotions. On MSP-IMPROV Target, ANGER is recognized most accurately, followed by NEUTRAL speech. Overall NEUTRAL utterances for all datasets are more accurately recognized than the combined class of emotional speech. Further, there is a high variation in performance between different emotions. Improving performance while focusing on poorly performing emotions like sadness and fear, which have an extremely bad performance, will help improve speech recognition.

| Corpus | Pearson | Spearman |
|---|---|---|
| CREMA-D | 0.73 | 0.86 |
| RAVDESS | 0.82 | 0.93 |
| MSP-IMPROV Target | 0.74 | 0.84 |

Table 2: Spearman and Pearson Correlation between the cross-comparison WER and the observed WER

Performance varies largely with sentences—some show excellent performance while others do not. Performance also varies across APIs—sentences with good performance with one system may perform well with others.

RAVDESS has two similar sentences, and hence it does not make sense to look at performance variation with spoken content on RAVDESS. On MSP-IMPROV Target, the WER varies from lower than 5% for some sentences to above 30% for others. Similar variations are also observed with Speaker Identity.

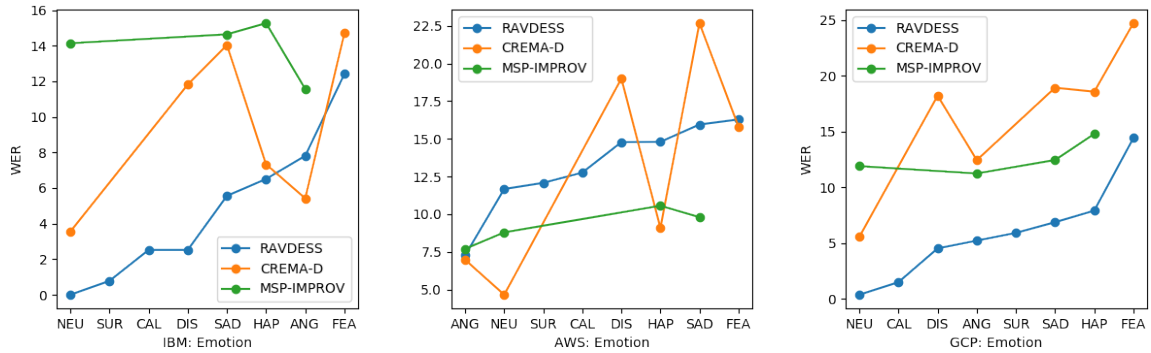
| Corpus | IBM AWS | IBM GCP | AWS GCP |
|---|---|---|---|
| CREMA-D | 0.84 | 0.74 | 0.85 |
| RAVDESS | 0.84 | 0.63 | 0.84 |
| MSP-IMPROV Target | 0.67 | 0.69 | 0.65 |

Table 3: Spearman Correlation between two-system cross comparison WER and the observed WER

### 5.2 Evaluating Performance without Manual Transcripts

When manual transcripts are not available, we treat the output of one API as the reference and get WER for other APIs with respect to it. We refer to this as the *cross-comparison* WER. We then change the reference API and repeat the process. The performance is reported as the average of all cross-reference WERs. In our case, we use three API's. Thus the average cross-comparison WER is the average of 9 cross-comparison WERs.

We test our metric on CREMA-D, RAVDESS, and MSP-IMPROV Target by computing Spearman and Pearson's correlations between the true WER and the average of the nine cross-comparison WERs. The correlations, mentioned in Table 2, are high, all above 0.7, indicating that the approximation is not perfect but overall accurate.




(a) IBM: WER on CREMA-D, RAVDESS and MSP-IMPROV Target (b) AWS: WER on CREMA-D, RAVDESS and MSP-IMPROV Target (c) GCP: WER on CREMA-D, RAVDESS and MSP-IMPROV Target

Figure 1: Performance based on emotion

| Dataset | Improvisation | F-value/ P-value | | | Improvisation |
|---|---|---|---|---|---|
| | | Sentence | Emotion | Actor | |
| CREMA-D | | 2163 | 56.6 | 5.73 | |
| RAVDESS | | 0.02/0.87 | 12.96 | 12.54 | |
| MSP-IMPROV Target | Complete | 71.78 | 6.99 | 13.37 | 32.42 |
| | Improvised | 42.89 | 5.92 | 5.58 | |
| | Unimprovised | 48.37 | 2.78/0.04 | 18.75 | |
| MSP-IMPROV Dialogue | | 0.11/0.73 | 0.81/0.48 | 2.43 | |

Table 4: Statistical significance of various factors on speech recognition performance. For entries where the P-value is not mentioned, it is almost zero.

We also conduct experiments to check whether the metric can be based on two systems instead of three. The Spearman Correlations are tabulated in Table 3. The two system cross-comparison metric is representative of the WER but the correlation is not as strong.

In Table 5, we report cross-comparison WER on the MSP-IMPROV Dialogue subcorpus which includes conversations used to evoke the desired emotion for the target sentence so that the required emotion sounds natural rather than acted. Each conversation is between a male and a female speaker. Emotions in natural speech are not as intense as they are in the acted versions. Also, only parts of the conversations contain emotional speech (neutral speech with parts of emotional speech) and it is natural to expect that the influence on recognition rates will be attenuated. It is however still present: ANGER and HAPPINESS have much worse recognition than NEUTRAL speech. Here however the best recognition is for SAD speech.

| Emotion | Cross-Comparison WER |
|---|---|
| SAD | 18.84 |
| NEU | 20.73 |
| ANG | 23.29 |
| HAP | 23.47 |
| **Overall** | 21.45 |

Table 5: Cross-comparison WER for MSP-IMRPOV Dialogue

## 6 Statistical Significance of Emotion, Speaker and Spoken Content

We now report the results of ANOVA analysis on each of the datasets, to compare the statistical significance of emotion, speaker and spoken content (sentence identity in our case) on performance. We compute the WER for each sentence separately. The F-values and P-values of the above-mentioned factors are listed in Table 4. As expected, spoken content has the highest impact on performance, other than RAVDESS which is expected to have low F-value for spoken-content. On CREMA-D and RAVDESS, Emotion impacts performance more than Speaker Identity.

On CREMA-D, the F-value for spoken content is about 40 times that of Emotion. Nevertheless, the Emotion and Actor Identity factors are statistically significant. Emotion has a much larger impact than actor identity. On RAVDESS, Emotion is slightly more impactful than Speaker Identity.

On MSP-IMPROV Target, the impact of Speaker Identity is more pronounced. However, on splitting the corpus based on whether the samples were improvised or not, on improvised speech, Emotion has a higher impact than actor identity. For non-improvised speech, Speaker Identity becomes important. Recognizing improvised speech, which is closer to natural speech, is more difficult. The impact of Actor Identity is thus lower (in improvised speech) than non-improvised speech. Note that the WER is higher for improvised speech (13.5%) compared to non-improvised speech (10.4%). Surprisingly, the impact of Emotion is higher in improvised speech.

For dialogues, Spoken Content has low significance, likely because factors are averaged out. Actor identity and gender of the interlocutor (together) impact recognition most.

## 7 Conclusions

We quantified the impact of Emotion on speech recognition performance. We developed a metric to analyze performance for audio samples where manual transcripts are unavailable and showed empirically that this metric works. In future work, we plan to analyze whether acoustic features are predictive of what sentences are likely to be misrecognized and the characteristic features per emotion.

## References


Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Ninth European Conference on Speech Communication and Technology*.

Theologos Athanaselis, Stelios Bakamidis, Ioannis Dologlou, Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. 2005. ASR for emotional speech: Clarifying the issues and enhancing performance. *Neural Networks*, 18(4):437–444.

Mohamed Benzeghiba, Renato de Mori, Olivier Deroo, Stéphane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, Richard Rose, Vivek Tyagi, and Christian Wellekens. 2007. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786.

Sahar E. Bou-Ghazale and John H. L. Hansen. 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans. Speech and Audio Processing*, 8(4):429–442.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed Abdel-Wahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affective Computing*, 8(1):67–80.

William Byrne, David S. Doermann, Martin Franz, Samuel Gustman, Jan Hajic, Douglas W. Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. Speech and Audio Processing*, 12(4):420–435.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "hey google is it ok if i eat you?": Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*, IDC '17, pages 595–600. ACM.

Libby Ferland, Ziwei Li, Shridhar Sukhani, Joan Zheng, Luyang Zhao, and Maria Gini. 2018. Assistive ai for coping with memory loss.

Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

John H. L. Hansen. 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, 20(1-2):151–173.

John H. L. Hansen and Sanjay A. Patil. 2007. Speech under stress: Analysis, modeling and recognition. In *Speaker Classification*.

Yusuke Ijima, Makoto Tachibana, Takashi Nose, and Takao Kobayashi. 2009. Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 4157–4160.

Theodoros Kostoulas, Iosif Mporas, Todor Ganchev, and Nikos Fakotakis. 2008. The effect of emotional speech on a smart-home application. In *New Frontiers in Applied Artificial Intelligence, 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2008, Wroclaw, Poland, June 18-20, 2008, Proceedings*, pages 305–310.

Hong Meng, Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. 2007. Combined speech-emotion recognition for spoken human-computer interfaces. In *2007 IEEE International Conference on Signal Processing and Communications*, pages 1179–1182. IEEE.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

YC Pan, MX Xu, LQ Liu, and PF Jia. 2006. Emotion-detecting based model selection for emotional speech recognition. In *The Proceedings of the Multiconference on" Computational Engineering in Systems Applications"*, volume 2, pages 2169–2172. IEEE.

Thomas S. Polzin and Er Waibel. 1998. Pronunciation variations in emotional speech. In *In: Proc. of the ESCA Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 103–1008.

Arsénio Reis, Dennis Paulino, Hugo Paredes, and João Barroso. 2017. Using intelligent personal assistants to strengthen the elderlies social bonds. In *International Conference on Universal Access in Human-Computer Interaction*, pages 593–602. Springer.

Mansour Sheikhan, Davood Gharavian, and Farhad Ashoftedel. 2012. Using DTW neural-based MFCC warping to improve emotional speech recognition. *Neural Computing and Applications*, 21(7):1765–1773.

Matthew A Siegler and Richard M Stern. 1995. On the effects of speech rate in large vocabulary speech recognition systems. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 612–615. IEEE.

Herman J. M. Steeneken and John H. L. Hansen. 1999. Speech under stress conditions: overview of the effect on speech production and on system performance. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999*, pages 2079–2082.

Frank A. Russo Steven R. Livingstone1. 2018. 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): e0196391.

Yanqing Sun, Yu Zhou, Qingwei Zhao, and Yonghong Yan. 2009. Acoustic feature optimization for emotion affected speech recognition. *2009 International Conference on Information Engineering and Computer Science*, pages 1–4.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.

David R. Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 2015. Evaluating spoken dialogue processing for time-offset interaction. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 199–208.