Classification of Whispered Speech Versus Normal Speech

Dominique Geissler (s1753053), Jelle Hamoen (s2189488), Sara Polak (s2377047)

University of Twente, the Netherlands

d.m.geissler@student.utwente.nl, j.g.hamoen@student.utwente.nl, s.polak@student.utwente.nl

Abstract

The adoption of speech processing systems is increasing and as such, more specific use cases are becoming apparent, one being the use of whispered speech. Since there is still a lack of applied whispered speech recognition research, the current project has focused on distinguishing whispered from normal speech such that this classification can be used in everyday situations. Existing studies mainly focus on specific features related to energy, time, and frequency. To easily classify whispered and normal speech based on these findings, this project used spectrograms in which most of these features are embedded.

For the classification, SVM, kNN and Decision Tree classifiers were trained and tested. The data for the classification contained samples whose amplitude was normalised and samples where this wasn't the case. It was found that the SVM classifier outperformed the kNN and the Decision Tree classifier. Since speed is a relevant aspect in real-life applications, future work could look into other features or classifiers to be used for classification, while keeping in mind the data storage that is required. After that, studies should also be conducted in real-life situations to establish ecological validity of the classification of whispered versus normal speech.

Index Terms: speech classification, whispered speech, whisper, SVM, kNN, Decision Tree, speech processing.

1. Introduction

Within the realm of speech technology, a lot of different speaking styles have been investigated over time. Examples include the study of the effects of emotion [1], accents [2], or noisy environments [3] on (understanding) speech. This project focuses on the difference between a normal speaking style and one where whispered speech is being used.

The distinction between normal speech and whispered speech in itself can be of significant use already as the use of whispered speech often indicates that the speech data is sensitive, or at least not appropriate for everyone in the environment to be heard. Even if the understanding of the whispered speech is not perfect (yet), users can value a speech application more if it listens to their needs by lowering the volume when giving possible replies or even giving these in a textual format as it might contain sensitive data.

However, at this moment, mostly empirical studies exist that look into the differences between normal and whispered speech but the application of this knowledge in the real world seems to be limited. Little studies could be found that report on automatically recognizing whispered speech in real-life situations. Therefore, it has been investigated in this project how well a classifier can distinguish whispered speech from normal speech, where the end goal would be to use this classifier in an everyday application.

An example of an application where whispered speech recognition is beneficial, is transcription. According to [4], automatic speech recognition (ASR) shows better performance if it is trained on a specific speech type. A classifier that recognises whispered speech could help the ASR to decide for the proper algorithm. A practical use-scenario for this would be the transcription of police calls, where the victim has to whisper so their culprit doesn't hear them such as in domestic violence or robbery situations. Another example would be a situation in which you need to share private information with a voice assistant, such as health records or credit card information which you do not want to share with everyone around you if you are in a public space. Then it would be good if the voice assistant whispers back to the user.

It is for these situations that whispered speech recognition can be used to improve people's lives. Therefore, the goal of this project was to find out where we stand with regard to the possibilities of differentiating normal speech from whispered speech. To this end, the following research question was investigated: **"How and to what extent can a classifier distinguish whisper from regular speech?"**

2. Related work

Different studies have been conducted on the topic of whispered speech versus normal speech as well as on the possible classification of these speaking styles. In 2002, a study was done on the classification of normal speech, whispered speech and softly spoken speech [5]. The study showed that normally phonated and whispered speech show differences in formant characteristics, i.e. an increase in formant frequency F1 for whispered speech as well as a smaller amplitude for this first formant F1. Moreover, it was noted by the authors that whispered speech has less energy (also related to amplitude) at the lower frequencies. These are all indications for features that can be used to distinguish whispered speech from normal speech.

A study by Zhang and Hansen provides an even more extensive classification of speaking modes, namely whispered, soft, neutral, loud, and shouted speech [6]. They investigated different features that could help to classify speech in these categories, including duration and silence percentage, frame energy distribution, and spectral tilt. Here, whispered speech is defined as the lowest vocal mode of speech with limited vocal fold vibration. It was found that sentence duration increases in all four speech modes relative to the neutral mode, which was mostly caused by silence duration increases for whispered speech (as opposed to increases in word duration). Next to this, the amount of low energy frames, i.e. silence and fricative

frames, is much more in whispered speech than in the other speech modes. Lastly, the decrease in spectral slope that was found for whispered speech implies that the energy in high frequencies increases when speech intensity decreases. Overall, the study showed that out of all different speech modes, whispered speech has the most discriminative characteristics. This can be seen as a positive reinforcement to try and classify the speech mode of whisper specifically.

More recently, Grozdic and Jovicic [4] noted that the performance of neutral-trained ASR systems significantly degrades when whisper is applied, while there are actually many circumstances in which whispered speech should be recognizable. They demonstrate that the use of deep denoising autoencoder (DDAE) and Teager Energy Cepstral Coefficients (TECC) can increase whispered speech recognition. Moreover, they noticed that there is a lack of corpora suitable for whispered speech recognition, so it was decided to create our own corpus for this project. Also, in the paper by Grozdic and Jovicic, it is shown how spectrograms differ between normal and whispered speech, especially for vowels, which makes it a potential factor to use in the classification process.

Another study by Fan, Godin, and Hansen analyzed not just the features that discriminate whispered speech from normal speech, but also the dependency of differences between normal and whispered speech on speakers and phonemes [7]. Fan, Godin, and Hansen define whispered speech as the absence of periodic vocal fold vibration in the production of phonemes that otherwise include such vibration. Since individual differences in producing (whispered) speech can cause great variations in speech recogniser performance, this is an interesting study to take into account. Not only a working classification is required but also a classification that still works when different speakers are detected, such that a robust system is created. Fan, Godin, and Hansen found that for vowels, the difference between whispered and normal speech was found to be consistent across speakers, especially above 4000 Hz. This led us to the decision of recording speakers with different accents and different genders as well.

As the studies discussed above show, there are, among others, enough features related to frequency, energy, and time that can help to classify whispered from regular speech. Therefore, spectrograms have been used in this project to try and classify these different speaking styles, as they include the mentioned features of frequency, energy, and time.

3. Method

For the recording process, seven participants, three male and four female, were invited. The recording process took place in a closed room in the library of the University of Twente. After the participants had received an explanation of the study, read the information brochure and signed the consent form, they were set up with a microphone. A lavalier microphone, part of a Sennheiser wireless system, was used to ensure proper recording, also of the softer whispered speech. Moreover, a Focusrite Scarlett 18i8 (second generation) audio interface was used during the recording process.

The participants were instructed to read two sets of 50 sentences, chosen from a set of sentences which is a recommended practice for speech quality measurements in [8]. Before each set, participants were requested to read out five practice sentences. These were used to make the participants feel more comfortable as well as to test whether everything was recorded properly. For the first set of 50 sentences, participants were asked to read them out in a normal voice (with a small pause between the sentences). For the second set of 50 sentences, the participants were asked to read them out in a whispered voice. Here, the practice sentences were also used to check whether participants would actually whisper or just lower their voice (but still producing voiced sounds).

Recorded sentences were stored in separate .wav files, which resulted in a data set of 700 sentences in total. The data was downsampled to 16.000 samples per second, to decrease the amount of data. Using Zenodo's Librosa [9], a Python package for music and audio analysis, a spectrogram was extracted for each data point. The spectrogram as a feature was used instead of the commonly used MFCCs, because whispering influences the source and not the filter. Not the articulators are different in whispering, but the vocal fold vibrations differ. As MFCCs model the filter, they are not the best feature to extract for this project. The spectrogram, in return, contains information about the source and is thus the better feature to use to classify between normal and whispered speech. Since the lengths of the wav. files varied, the spectrogram data were filled with zeros until they all had the same length. The processed spectrogram data and their respective label were stored in SciPy's [10] Pandas [11] dataframe with the rows representing the data points and the columns representing the spectrogram data. The data was then split into a training and a test set with a 80-20 ratio, respectively.

For the classification, initially two classifiers were trained, Scikit-learn's [12] k-Nearest Neighbor (kNN) and Support Vector Machine (SVM). kNN looks at a data point and calculates the k-nearest neighbors. It then assigns a label to the new data point based on the label that is most common among its neighbors. The default settings of kNN were used: k is 5 and the weights were uniform. SVM constructs a hyperplane in a multidimensional space that separates the classes. It does this by iteratively looping over the data to find the best separation of the classes. Also here, default settings were used.

The classifiers were trained on random 80% of the dataset and subsequently tested on the remaining 20%. After training and analysing the classifiers, it was decided to normalize the amplitude of the data, to ensure that classification is not purely based on the amplitude of the recordings. In a real-world use case, the distance between the speaker and the microphone is not going to be the same, and as such the amplitude of the incoming signals would not be constant. This could hypothetically mean that whispered speech close to the microphone could yield similar amplitudes to normal speech further away.

After normalisation, the kNN and SVM classifier were retrained on the normalized data points. In addition, it was decided to also train and test a Scikit-learn Decision Tree (see Appendix 1). This was done to be able to compare the performance of different classifiers. Moreover, the Decision Tree can be used to estimate which features were of importance in classification of the normalized set.

4. Results

Using the Scikit-learn classification report function, the performance of the trained classifiers was assessed. This yielded the following results:

Table 1: Results from the classification

	kNN raw	kNN norm	SVM raw	SVM norm	Decision Tree norm
Precision (whisper)	0.96	0.69	1.00	0.98	0.90
Precision (normal)	0.99	0.96	1.00	0.99	0.95
Recall (whisper)	0.99	0.97	1.00	0.98	0.93
Recall (normal)	0.96	0.66	1.00	0.99	0.92
F1-score (whisper)	0.97	0.80	1.00	0.98	0.92
F1-score (normal)	0.97	0.78	1.00	0.99	0.94
accuracy	0.97	0.79	1.00	0.99	0.93

SciPy's classification report gives back the precision, recall, F1 score and accuracy. Since the recognition of either class is just as important, we strive for an optimal balance between precision and recall. Thus, the evaluation of the performance of the classifiers is assessed using F1 and accuracy.

When looking at the results of the classification of the raw data, it can quickly be seen that the performance of the kNN and the SVM classifiers was very high. SVM had a perfect score, with 100% accuracy and an F1 score of 1.00 for both classes. kNN performed slightly worse with an F1 of 0.97 for both classes and an overall accuracy of 97%. After normalisation, the performance of the classifiers decreased. For the kNN classifier, the accuracy went down by nearly 20% to 79%. Also, the F1 dropped to 0.78 and 0.8 for normal and whispered speech respectively. This indicates a sharp decrease in performance for the kNN classifier. When it comes to the SVM, performance also decreased, but not as drastically as with the kNN. The accuracy went down slightly to 99% and the F1 for normal and whispered speech were 0.99 and 0.98, respectively. The performance is thus still very high. Scikit-learn's Decision Tree has a high performance as well. With an accuracy of 93% and an F1 score of 0.94 and 0.92 for normal and whispered speech respectively, the Decision Tree performs nearly as well as the SVM. Overall, the SVM outperforms the kNN and the Decision Tree.

5. Discussion

All in all, the classifiers performed well in recognizing normal and whispered speech with accuracies mostly over 90% and high F1 scores. It can be seen that performance is higher when the amplitude is not normalized, especially in the case of the kNN classifier. For unnormalised data, the kNN classifier performed nearly perfectly, but for normalised data performance dropped drastically with the accuracy going down to 79% and the F1 scores dropping to roughly 0.8 for each class. We believe that the imbalance of F1 between the normalised and raw kNN classifier can be attributed to the nature of the classification method. kNN attributes a label to a data point based on the most common class among its neighbours. When normalising the data, the data points come closer together, leading to the classifier being unable to make as precise predictions. The imbalance in performance of the kNN could also be interpreted as an indication of mislabeled data. When one side of the discrimination vastly outperforms another one it can be an indication of mislabeled data. However, as the other methods use the same dataset and are much more balanced this notion can be rejected.

In general, SVM outperforms kNN and Decision Tree. The SVM on raw data has an accuracy of 100% and an F1 score of 1.00 per class. This could be an indication of the classification task being too easy for the classifier or that there is something wrong with the model: the dataset might be insufficiently large to warrant similar results in context. In an attempt to reduce those concerns the effort was made to improve generalisation by normalising the audio data before use in the classifiers. After normalisation, the performance of SVM was still very high with an accuracy of 99% and an F1 score of 0.99 and 0.98 for normal and whispered speech respectively. This shows that it is slightly more difficult for SVM to recognise normalised whispered from normal speech.

The Decision Tree classifier performed nearly as well as the SVM classifier on the normalised data and a lot better than the kNN. This shows that the Decision Tree is able to classify normal and whispered speech in an effective manner. Overall, SVM performs best at the classification task. Even though significant results were expected (F1 around 0.7), results were significantly better than initially anticipated.

Since SVM performs so well at classifying whispered speech, it might be a good choice for real-life applications. For a transcription application, the SVM could be very reliable in recognising whispered speech and thus activating the right, specialised ASR. This can increase transcription quality. For voice assistants, reliable whispered speech recognition can lead to more desirable behaviour.

While the results are promising, their representativeness for real-life applications can still be questioned. The recordings took place in a quiet lab using a lavalier microphone. This ensured good data quality, but is not representative for real-life situations, as there might be background noises or the user is not so close to the microphone. Thus, the classifier would then have to deal with a lot of noise or bad data quality, making the classification task harder.

Also the sampling rate was still rather high, with 16.000 samples per second. This led to a potentially unnecessary amount of data, which less powerful processors such as mobile phones might not be able to store and process. In order for this classifier to be efficient in terms of space and time, the sampling rate might need to be lower. Another factor that leads to a rather large amount of data is the use of the spectrogram as a feature. The spectrogram represents the amplitude, the frequency and the time domain of the waveform, leading to a lot of data. Other features, such as MFCCs, might lead to a smaller amount of data. This would require less storage and could speed up

computation. At the moment, the computation time is rather high, which is insufficient for real time applications. Transcription applications or voice assistants require classifiers with little computation time as they need to act fast. At the current state, the classifier cannot fulfil these requirements yet. In addition, the data structure can be questioned. For the classification, all the data was stored in one dataframe with 700 rows and roughly 113.000 columns. This is a big data frame and it might be easier to get an overview of the data if each data point had its own dataframe. This would enable us to see which data the classifier uses for classification more easily.

6. Conclusion

This research focused on the question: "How and to what extent can a classifier distinguish whisper from regular speech?" To answer this question, a group of seven people of mixed gender was recorded, with three male and four female speakers. The participants were asked to read 100 sentences out loud of which 50 were to be spoken normally and 50 were to be whispered. Using a lavalier microphone, the recordings were of high data quality. These recordings were sampled down to a sampling rate of 16.000 Hz to decrease the data amount. For the classification, a spectrogram per sentence was extracted. These data points were classified using kNN and SVM classifiers. The performance of the classifiers was very high with SVM outperforming kNN. As amplitudes can vary greatly in real-life applications due to the speaker being closer or further away from the microphone, the amplitude of the data samples was subsequently normalized. kNN and SVM were trained and tested again. On top of that, a Decision Tree classifier was used to gain insight into the features used for classification. Also on a normalized data set, SVM outperforms kNN and Decision Tree.

Future work could look into using other features for the classification, while keeping in mind the storage space the data requires. There are probably features that require less storage space and still deliver satisfactory results. Less data also leads to faster computations. In addition, it might be good to look into different normalisation methods besides amplitude normalisation as that might have an effect on the performance of the classifier. Moreover, future work could look into using other classifiers. Some classifiers like neural networks require a lot of data, others can deal well with little data. Future research could try to find the best classifier for a certain amount of data input. Additionally, future work could try classification using more realistic data with background noises and differing proximity to the microphone. Lastly, classification could be tested in real-life situations with users to see whether the application can deal with this challenge. This would also expose the classifiers to speakers which were not present in the training set.

7. Acknowledgements

We thank our supervisors for the continued support during this project and our participants for the effort they put into the recordings.

8. References

[1] C. E. Williams and K. N. Stevens, "Emotions and Speech: Some Acoustical Correlates," *J. Acoust. Soc. Am.,* vol. 52, no. 4B, pp. 1238-1250, 1972.

[2] A. D. Lawson, D. M. Harris, and J. J. Grieco, "Effect of foreign accent on speech recognition in the NATO N-4 corpus," *8th Eur. Conf. Speech Commun. Technol.,* pp. 1505-1508, 2003.

[3] T. Shimizu, K. Makishima, M. Yoshida, and H. Yamagishi, "Effect of background noise on perception of English speech for Japanese listeners," *Auris Nasus Larynx,* vol. 29, no. 2, pp. 121-125, 2002.

[4] D. T. Grozdic and S. T. Jovicic, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering," *IEEE/ACM Trans. Audio Speech Lang. Process.,* vol. 25, no. 12, pp. 2313-2322, 2017

[5] S. J. Wenndt, E. J. Cupples, and R. M. Floyd, "A study on the classification of whispered and normally phonated speech," *7th Int. Conf. Spok. Lang. Process. ICSLP 2002,* pp. 649-652, 2002.

[6] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," *Int. Speech Commun. Assoc. - 8th Annu. Conf. Int. Speech Commun. Assoc. Interspeech 2007,* vol. 4, pp. 2396-2399, 2007.

[7] X. Fan, K. W. Godin, and J. H. L. Hansen, "Acoustic analysis of whispered speech for phoneme and speaker dependency," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH,* no. August, pp. 181-184, 2011.

[8] "Harvard Sentences." [Online]. Available: https://www.cs.columbia.edu/~hgs/audio/harvard.html. [Accessed: 14-Jan-2020].

[9] Brian McFee, Vincent Lostanlen, Matt McVicar, Alexandros Metsai, Stefan Balke, Carl Thomé, Colin Raffel, Ayoub Malek, Dana Lee, Frank Zalkow, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Ryuichi Yamamoto, Scott Seyfarth, Eric Battenberg, Виктор Морозов, Rachel Bittner, Keunwoo Choi, Josh Moore, Ziyao Wei, Shunsuke Hidaka, nullmightybofo, Pius Friesch, Fabian-Robert Stöter, Darío Hereñú, Taewoon Kim, Matt Vollrath, and Adam Weiss, "librosa/librosa: 0.7.2". Zenodo, 13-Jan-2020.

[10] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2019) SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. preprint arXiv:1907.10121

[11] Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)

9. Additional material

Appendix 1: The Decision Tree output