Don't over think,
I know your drink

Suzanne Bardelmeijer, Karen Beckers,
Sanne Eggengoor, Job van Gerwen & Roos Slingerland

The front page image is cover art from Marshall McLuhan

Don't over think, I know your drink

Suzanne Bardelmeijer 10716971
Karen Beckers 10811958
Sanne Eggengoor 10729895
Job van Gerwen 10378863
Roos Slingerland 10775935

Project Report
Course Media Understanding
Credits: 6 EC

Bachelor Artificial Intelligence

College of Science
University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Course coördinator
Dr. A. Visser

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

April 2th, 2017

1 Summary

This research developed a system Sjeel 2.0 that can recognise the personal wishes of the two favourite drinks of students: beer and coffee. It uses two different media types namely speech recognition and speaker identification. Using Google's API for speech recognition the difference between coffee and beer a result was reached of only 3 mistakes in 199 samples. Using multiple features such as pitch, MFCC and RMSE for speaker identification multiple algorithms with 10-fold crossvalidation were applied. Naive Bayes, Decision Tree, Random Forest all reached an accuracy around 70% while Gradient Boosted Trees after parameter finetuning reached an accuracy of 81%. As expected, the level of noise does influence the accuracy of the system. When recording the trainingsset in a quiet environment, an accuracy of 95% was reached. This system could be used in situations where speaking for a long time is an issue, such as with stutters or emergency situations. Future work could focus on extending the number of beverages and users. Also other media forms for speaker identification could be added such as face recognition and user-dependent vocabulary.

Contents

2 Introduction

'Goodmorning, a large cappuccino I assume?', asks Sjeel. The woman of the canteen of Science Park knows the preferences of her regular customers almost flawless. How great would it be to copy the skills of Sjeel and use them in a system? A system that can recognise the preferences of the users by analysing their voices. And above that, can recognise the kind of beverage you want. To achieve this, 2 different kinds of media are used: speech recognition and speaker identification. This would not only save time, but is also a great solution for people who have troubles with speech.
Imagine a residential care home with many people requiring help. How efficient would it be that in case of emergency a person in help only has to say 'help' and our system would know who it is. And above that, knows the medication history and can therefore provide help more efficiently.
Another example is with people suffering from stuttering. Instead of saying: *I would like to have a cappuccino with almond milk and a dash of cacao*, say: *Coffee please*. This would save even more time then it would with regular speakers. Secondly, it might decrease the social anxiety of speaking in a busy queue under pressure.
Even though this system could be used very widely, there is decided to define the two-folded problem as follows. First, the system recognises which beverage is ordered by using speech recognition. Secondly, by using speaker identification, the system analyses the personal preferences of this specific user. In this study the number of users is equal to the number of groupmembers, so 5 in total.
For example if a user says: 'Beer please', the system should know that the user ordered the beverage beer. Secondly it should recognise that it was user 3 that ordered the beer and would therefore know that it should be a Heineken. The beverages coffee and beer are chosen because they are widely known as the most popular student drinks.
The system is expected to prove the following hypothesis:

1. Distinct the word beer from coffee perfectly.
2. Distinct the man of the group from the 4 women perfectly.
3. Reach an accuracy of at least 80% for recognising the speaker in a noisy environment.
4. Reach a higher accuracy for recognising the speaker in a quiet environment.

This research tends to bridge the semantic gap by giving meaning to the values of audio signals. What to a computer just a collection of wavelengths is, is to humans the difference between 'beer' and 'coffee' or between the voice of person A and person B.

3 Theoretical Foundation

3.1 Phonetics

To understand the conversion from pure sound waveform to text and to the speaker identity, it is necessary to have a better understanding of speech itself.

In this section a short introduction in the field of phonetics is given. In a strict sense, speech is produced by the vocal tract, which is a collaboration between the lungs, larynx and oral cavities. The larynx contains the vocal cords, two small pieces of tissue that vibrate when they are closed and air is pressed through. But with this only the pitch of the voice can be altered. Different sounds can be made by changing the shape of the oral cavities. For example, the difference between an [i] in 'keen' and [o] in 'hope' can be explained by the change in the location of the tongue, the height of the lower jawbone and the tension in the lips. Also the difference between [m] and [n] in 'man' can be explained by the difference in location of the tongue and lips. The difference in sound comes from the resonance in the cavities.

This is of course also visible in the waveform of the sound, which is a combination of the frequency of the vocal cords, and the frequency of resonances in the several cavities. These frequencies are called formants, with the lowest formant (formant 0, f_0) as the frequency of the vocal cords. The next two formants contain information about the resonance in the pharynx (cavity in the throat) and mouth. These formants are easily changeable by moving the tongue and jaw and thus contain the most information about what is said. The next formant describes the resonance in the nasal cavity and is sometimes used to identify nasal consonants (such as [m] and [n]). The fourth and fifth formant contain speaker specific information, because this describes the resonance in other cavities, which cannot be changed when speaking. Therefore the first, second and third formant are the most useful in defining what is said and the fundamental frequency (f_0) and the fourth and fifth formant contain the most useful information about who spoke.

3.2 Speech recognition

There is an overall agreement in the literature that Hidden Markov Models (HMM) is the best approach for Automatic Speech Recognition (ASR) (Baker et al., 2009). In almost all applications a HMM forms the basis of the system. The models are trained for each word in a large dataset that contains labeled speech data. When a new (non-labeled) input is given, the sentence is recognised by the most likely model sequence (Foote, 1999). For large-vocabulary recognition datasets a different approach is applied. Instead of making a HMM for every of the thousands of words in the dataset, a much smaller number of sub-word models are constructed that can be combined to form a word model. Since the that will be build in this research only has to recognise two keywords, "beer" and "coffee", the first approach will probably be best for this problem.
Even though HHM is the standard method for building these kind of systems, new research focusses on deep neural networks. Google has build a neural network based keyword spotting system that is 45% more accurate than Hidden Markov Models (Chen, Parada, & Heigold, 2014).

3.3 Speaker identification

As stated in the section about Phonetics, the most useful data in speech are the fourth and fifth formants and the fundamental frequency. To be able to obtain these data from the waveform sounds there needs to be done a feature extraction (Kinnunen & Li, 2010). These features are typically the Mel Frequency Cepstral

Coefficients for the previous 20-30 ms. However, it might be possible to use only the pitch of the voices, because the choice is limited to 5 speakers (Weenink, 2017).

Mel Frequency Cepstral Coefficients (MFCC) are based on the nonlinear and logarithmic perception of pitch in the human ear (Hasan, Jamil, Rahman, et al., 2004). The frequencies on the Mel scale are below 1000 Hz linearly spaced, above 1000 Hz the frequencies are logarithmically spaced. This has an analogy with the human ear, where the basilar membrane (located in the cochlea) vibrates at specific points (depending on the frequency of the sound). The spacing of these specific points is the same as the Mel Frequency Scale, which makes this Scale very suitable for speaker identification.

Guassian mixture models are widely used in unsupervised speaker recognition systems (Reynolds, F., & Dunn, 2000). The task of such a system is to determine by who a segment of speech was spoken. The system described by (Reynolds et al., 2000) was the *Gaussian Mixture Model-Universal Background Model* (GMM-UBM). This system is based on the optimal likelihood ratio test. *GMM-UBM* uses *Gaussian mixture models* for likelihood functions. For the representation of alternative speakers in a specific segment of speech a universal background model is used and Bayesian adaptation obtains a hypothesized speaker model.

3.4 Speech recognition and speaker identification combined

This research aims at simultaneously recognizing the person who is speaking and what is being said. For the system to work it is therefore necessary that the components mentioned above are combined. To this, though, there is little research done. In 2010 a self-learning speech controlled system was presented by Herbig, Gerl en Wolfgang. (Herbig, Gerl, & Wolfgang, 2010). Their research tried to improve speech recognition by first identifying the person who is speaking. Speaker identification with high accuracy was obtained by first retrieving personal speech characteristics such as the Mel Frequency Cepstral Coefficient. Speech recognition was improved by taking into account how the identified person speaks. Knowing how this person pronounces specific words or formulates sentences improves the recognition of what is being said.

This project will not focus on improving speech recognition, but uses the identification of a speaker to fulfill his or her needs by knowing what his or her preferences are.

4 Research Method

4.1 Data

Two datasets of both 200 audio files are created. One dataset was situated in a noisy environment and the other in a quiet environment. For each dataset all five users have recorded the dutch sentence "Mag ik koffie?" (Can I have a coffee?) and the sentence "Mag ik een biertje?" (Can I have a beer?) 20 times. A screencapture (Figure 5) of the created dataset can be found in Appendix A.

4.2 Feature extraction

To create an application that is able to define the preferences of a specific user, a Python program is created that can capture sound of the input using the *PyAudio* toolkit (Pham, 2017). This toolkit can store an audio input into a Waveform Audio File Format (WAV).
After the audio has been captured and stored, the program calls the *SpeechRecognition* library (Zang, 2015), which makes use of the *Google Speech Recognition API* in order to translate the audio into text. This library can be used for both direct audio input and previously recorded audio. The approach that uses previously recorded audio will be applied in the training phase and the approach using direct audio input will be applied in the test phase.
RMS Energy is a metric that measures the continuity of the audio signal level (Weihs, Jannach, Vatolkin, & Rudolph, 2016). A piece of audio that contains many portions of silence will have a large RMSE value whereas audio containing continuous signals will have a small RMS Energy value. The RMS Energy can be computed by taking the root-mean-square (RMS) energy of a signal (Weihs et al., 2016). This project used the RMS Energy feature to trim the audio segments because in this way the silences can be cut off. This results in an audio segment that only contains speech. The differences in RMS Energy between segments of speech of different users will also help by the identification of a speaker. This is because people tend to speak with different dynamics.

In order to identify the speaker, several timbre features will be extracted from the audio files. These features are going to be the data for the classification algorithm. In order to extract these features, the *LibROSA* package (McFee et al., 2017) is employed. This package is commonly used for music and audio analysis and music information retrieval, but is also for this research a helpful tool, since it enables the extraction of a large number of features that can be used for speaker identification. As mentioned before the interview with an expert in the field of linguistics revealed that the most useful features for this research will be pitch and the Mel-Frequency Cepstral Coefficients (Weenink, 2017). These will be the features used for speaker identification.

4.3 Model training

The next step is training the algorithm used for identifying the speaker. Only supervised learning algorithms will be tested and a comparison between these algorithms will be made. The literature showed that Naive Bayes classification is commonly used in this area of research (Reynolds et al., 2000). This algorithm will therefore be applied in this study. Besides that, also a Decision Tree, Random Forest and Gradient Boosted Trees algorithm will be tested.

The Naive Bayes algorithm is based on Bayes theorem. Here the assumption of independence between every pair of features is made (Reynolds et al., 2000).

Decision Trees build a tree-like structure based on the characteristics of the data. Every node is a feature and the branches are the different values that a feature can have (Mitchell, 1997).

Random Forest learns multiple Decision Tree models simultaneously. The output is the most frequent label of all the predicted values of the models. Hereby the problem of overfitting that Decision Trees are prone to is solved

(Breiman, 2001).

Gradient Boosted Trees start with a pretty weak model and then tries to enhance that. It calculates the error of the model and fits a new Decision tree to the corresponding cost function. By adding this new model to the previous one, the whole function gets more and more complex. This happens in an iterative way in order to refine the model (Friedman, 2001).

4.4 Output: Sjeel in action

When both speech recognition and speaker identification have been executed, the results will be combined to evaluate what the speaker wants. The preferences of a person are predetermined by a questionnaire for a set of situations. The wish of the user will then be returned as a text message.
For this research the focus will solely be on detecting the preference for a beverage. This will be divided in two categories: beer and coffee. For both categories the users will be given a choice between several brands or flavours. Table 1 shows the preferences of the different users. The system will first tell the user that his/her preferred drink is on its way and will then show a photo of the user with either a cup of coffee or a beer in his/her hand. Figure 1 shows an example of the output.

User	Coffee preference	Beer preference
1 (Roos)	Glutenfree coffee	Glutenfree beer
2 (Karen)	Latte macchiato with soy cream	Alfa
3 (Sanne)	Black	Grolsch
4 (Job)	Cappuccino	Hertog Jan
5 (Suzanne)	Flat white	Duvel

Table 1: User preferences

Figure 1: Ouput example

4.5 Evaluation

Evaluation of the algorithms will be done by calculating the percentage of the correctly classified training examples. A 10-fold crossvalidation was applied to obtain reliable results.

5 Results

First of all the MFCCs of the audio files have been calculated. Examples of these MFCCs are in figure 2.

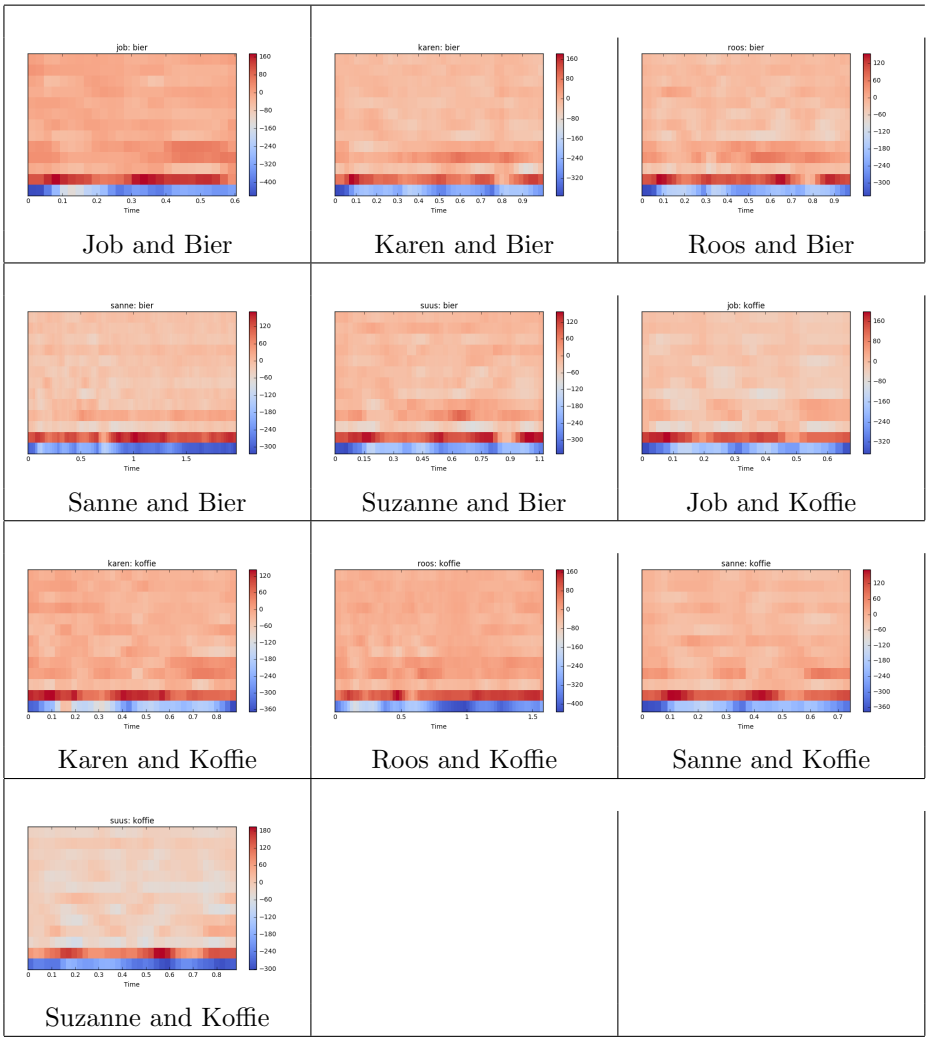

Figure 2: MFCCs of the speakers

Multiple learning algorithms have been used to obtain optimal results for both the quiet and the noisy environment. As shown in Figure 3 *Gradient Boosting Tree* yielded the best results in a noisy environment but *Naive Bayes*, *Decision Tree* and *Random Forest* have been tested as well. For the quiet environment, Figure 4 shows that Naive Bayes and Random Forest both yielded the same best result.

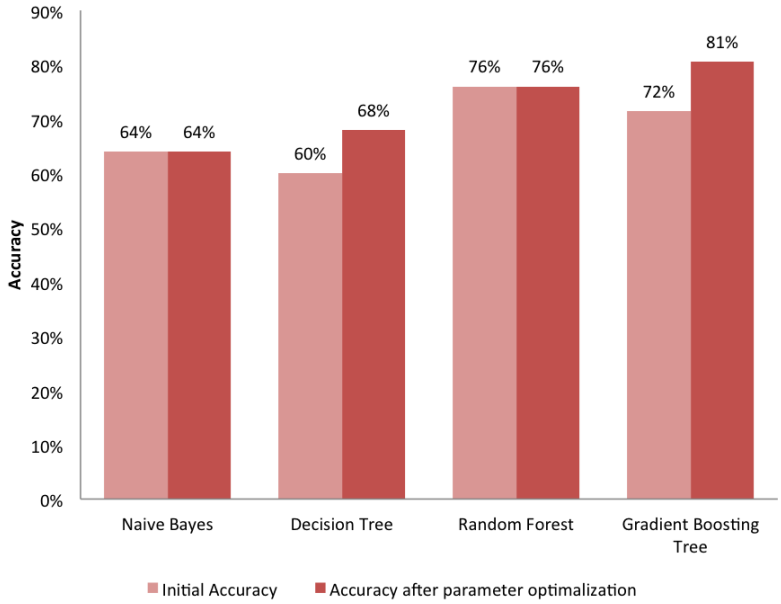

Figure 3: Obtained accuracies (crossvalidation 10) noisy environment

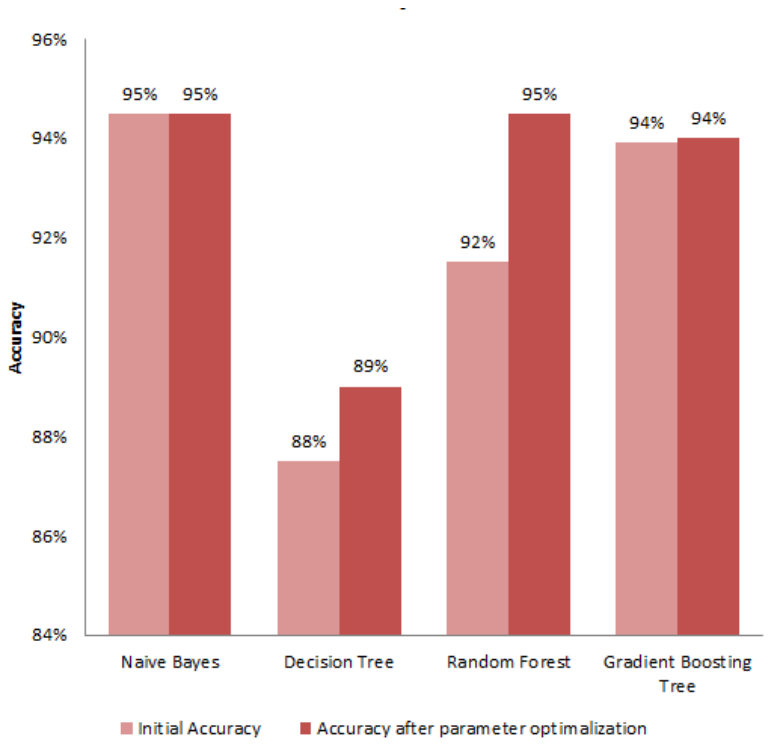

Figure 4: Obtained accuracies (crossvalidation 10) quiet environment

All tests made use of *10-fold-cross -validation.* An accuracy of 80,5% was obtained by using the *GradientBoostingClassifier* in Python. The corresponding parameter values after optimalisation are shown in Table 2.

Parameter	Value
n estimators	100
learning rate	0.5
max depth	1
random state	0

Table 2: Parameter values

6 Conclusion

In this paper a system was created that copied the skills of coffeelady Sjeel. By using the media speech recognition and speaker identification a twofolded problem was tackled. The system could both distinct between beverages as between users. When looking back at the hypothesis the following conclusions can be drawn.

1. *Distinct the word beer from coffee perfectly.* The system made only 3 mistakes in 199 descriptions, so this hypothesis was correct.

2. *Distinct the man of the group from the 4 women perfectly.* The system made no mistakes by distinguishing user 4 from the other users, so this hypothesis was correct.

3. *Reach an accuracy of at least 80% for recognising the speaker in a noisy environment.* The highest accuracy was reached after parameter finetuning of the Gradient Boosted Trees algorithm namely 81%, so this hypothesis is correct.

4. *Reach a higher accuracy for recognising the speaker in a quiet environment.* The highest accuracy was reached after parameter finetuning of the Gaussian Naive Bayes algorithm namely 95%, so this hypothesis is correct as well.

This research could in the future help people that struggle with speaking such as people who stutter, suffer from anxiety in social situations and in case of emergency where only a few words can be pronounced. Above that, it saves time of explaining all the details of your personal order.

7 Discussion

Looking back there are a few critical points to be made about this research.
First of all the group of users was small (only 5), for a real-world system this could be expanded for more users. The evaluation of how well the user was identified can then have more meaning.
Secondly, the number of different beverages could be expanded as well. Above drinks, other preferences could be added, such as medication or favorite songs.

When looking at the speaker identification other media could have been used as well. Visual recognition could have been an option as well. This could have been done by recognising the faces of even lengths or postures of the users.
This research now chose to let the users speak the exact same words. Literature shows that by using, for example, fillers, speaker identification can be improved. Future work can therefore focus on enlarging the number of users, beverages (or other preferences) and subjoin other user recognition techniques such as visual techniques and user-dependent vocabulary.

Appendix A

	Pitch	RMSE	MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13	User
22	-0.23	0.057498485752	-359.27475003	139.73980965	-30.4845936324	38.4005384131	3.6880076301	-49.3644062076	-12.3570046349	2.83108999328	-13.6192497426	-14.9058488941	-14.1465385834	-22.2099436239	-12.5822295791	job
184	-0.12	0.044482961297	-387.6973616636	132.9355508681	-17.1216728606	31.001299036	6.6747904017	-28.0888708635	-2.2439161825 4	-7.8875312264 2	-11.0645636392	-16.3170225059	-12.0330871731	-11.051587139	-16.9033655683	suus
130	0.22	0.065503954874	-361.2303208 26	130.2544753 91	-24.7159723515	44.8290236014	-2.9159236482 4	-36.8481411571	3.3225114051 8	0.9401362951 44	-23.4961893102	-13.8688210115	-10.9359778802	-10.2623613174	-7.4760072214	sanne
23	-0.45	0.058286424726 2	-342.3457752 83	139.7947472 57	-35.7106789655	35.4071778965	-0.6661545203 07	-49.9113713009	-16.5556179534	5.0844527636 7	-13.4835902073	-19.0028248676	-13.462344266	-23.0444046118	-12.8497057644	job
99	0.01	0.080877490737	-333.4931963 56	117.3991167 03	-42.1947962709	55.9360699041	11.9463092589	-40.2225025862	0.4445768454 87	-13.7092815703	-22.1845034806	-10.2694797456	-17.1995470431	-15.6327972601	-9.0180305043 8	roos
66	0.01	0.078055589666	-346.5048267 01	124.3118606 84	-12.7802886946	24.1720013764	-19.5947713042	-39.9882991084	-8.2026996746 6	2.1415315010 6	-9.3767668814 6	-29.2962232615	-18.2088684962	-1.2901393975	-26.2855907845	karen
172	0.1	0.072282053530 2	-330.0899430 21	146.2556474 77	-22.7055112592	63.4018672285	5.5858682555 51	-51.5925586955	-20.6120980165	-2.8442317244 6	-7.3979651643 7	-23.7863205144	-16.9482275 95	-10.3603541027	-13.3286367064	suus
141	-0.44	0.052689276636 6	-366.4857719 22	130.9315304 59	-4.1612337319 9	41.4411772441	-6.9942288448 4	-40.1317534778	9.3473224873 6	10.064899354 3	-28.0822374183	-25.8609162771	-9.8773301186	-17.972832331	-5.9494587347 2	sanne
165	-0.49	0.053229000419 4	-362.7309513 75	141.1777537 77	-25.6624713158	52.3381195498	12.1395596335	-46.3250433724	-22.9341524408	0.0683360768 03	1.6291177154 6	-23.8195332077	-21.4000326915	-10.979123787	-14.7063678449	suus
5	-0.05	0.069439813494 7	-294.8303490 95	146.4468828 75	-61.8047111011	31.2644796924	21.4848306024	-29.0701953637	-20.0464475615	-21.6641747784	-21.6865122141	-11.184721917	-13.2919194486	-24.1567782034	-12.8152933505	job
169	0.06	0.070533141439 8	-338.7105020 96	130.3914088 12	-34.8071255622	62.7186186096	0.7506317285 56	-41.6705493597	-6.6008345907 7	-14.2659026126	-17.4388157033	-12.386827996	-27.910134300 5	-14.7186963386	-3.3891641721 3	suus
16	-0.3	0.058947626550 13	-320.4960987 73	131.9727523 16	-66.1563129036	44.4247368346	22.4175718242	-40.6853550142	-17.3856058193	-17.5682293809	-16.6794586323	-8.1364176264	-12.884934662 6	-27.4968947277	-11.2643507344	job
88	-0.3	0.088332898914 8	-331.3219522 711	110.2349202 61	-19.1832520896	61.7817741792	-0.4083361494 59	-48.7571753889	-4.0213721232 6	-2.2735131128 7	-19.0628008596	-19.3792076017	-12.981494851 2	-10.6857279075	-10.3288217338	roos
18	-0.39	0.071893812259	-310.8688685 84	137.7570763 69	-62.2059213066	43.1217018971	27.5422535809	-38.6827023001	-18.2474082527	-15.6955147908	-24.3856035978	-12.7594309513	-10.273965896	-26.7352653541	-18.5608676047	job
20	-0.18	0.059874143455 15	-324.7338156 3	138.2818343 14	-32.5740710442	50.1483785294	-6.0705344366	-52.4608051556	-7.8686164995 5	-1.9570695543 4	-11.2095797 42	-21.943900822 1	-12.475347684 4	-17.8152419729	-19.9669029329	job
135	-0.37	0.064074568450 5	-365.9631527 97	119.1880435 04	-20.3493255039	50.8929832686	-8.0086564025 8	-35.0210693645	7.0057220034 7	-0.3675006632 13	-22.448838318	-16.1313460619	-16.059393492 9	-12.4131779361	-4.8238789057 3	sanne
175	-0.48	0.031745284795 8	-399.1643480 49	138.6023068 56	-53.0712174353	68.5564569099	28.4658206026	-54.8303370482	-32.6961937928	-18.1017166037	-6.8809688125 2	-15.6807521758	-17.735268879 7	-13.9413633481	-21.7011955582	suus
196	-0.36	0.083139523863 8	-330.7022823924	117.0348685 08	-21.9336717551	34.1808845079	4.0888331116 83	-29.0157871735	-7.4280936161 8	-15.3250981439	-20.0353966352	-18.2426812916	-10.313950254 5	-9.5140015175 4	-27.1637419452	suus
35	0.19	0.053117807954 5	-349.9620599 27	136.8270739 82	-31.9329102243	38.1826715305	-1.9071626504 2	-45.9714910365	-14.9094550857	2.2802930759 3	-10.9733737899	-18.3140983414	-14.622658736 1	-23.8496374246	-12.1374429052	job
59	-0.4	0.079629951265	-321.5607712 76	126.7886239 28	-28.501726274	45.7026658626	6.2272485945 5	-35.2916174822	-28.6164268432	-5.8341416772 2	-5.1268325574 7	-20.150033143	-17.527708507 4	-8.2191296536 8	-18.8178665214	karen
183	-0.42	0.045860704034 6	-394.6466499 77	134.0251111 91	-12.2013491286	32.9784740768	5.0181610424 1	-32.663604067	-12.9111091965	-3.2092517775	-14.0523021174	-22.876888106 5	-12.80057493 5	-11.185360887	-19.6823516138	suus
198	-0.19	0.082671768963 3	-335.7503463 76	132.1286279 96	-17.6660916221	38.2992803379	7.2318357496 6	-36.8798530 21	-12.3822085675 5	-9.5915629294	-24.3219473075	-17.764080358 9	-8.6798955708 2	-12.9914830417	-23.3786746343	suus
160	0.04	0.040016654878 9	-392.3989338 336	132.8953215 61	-24.8653983698	42.325146557	-6.1334768037 9	-34.4238543958	-10.0532602236	-16.6469217389	-17.3190389248	-19.555461286	-5.8462101257 5	-7.6039913762 8	-33.0850796883	suus
48	-0.33	0.077484026551 2	-313.3049546 82	131.5245546 72	-28.6713192821	49.8363378205	-4.3217424851	-42.9837332668	-18.151808842	1.8701595572 4	-5.6477911845 3	-30.696410388 9	-16.952821301 4	3.0179571801 1	-14.1294686172	karen
13	0.35	0.061743743177	-294.1007990 39	149.4285060 37	-67.2845666627	49.9641612395	18.702198343	-35.8674815045	-7.3457254369 5	-13.4101212932	-18.4747813602	-6.2007674626 2	-9.8665001177 9	-21.2139930084	-13.5207777889	job
174	-0.22	0.047209930048	-363.5783769 85	140.8471626 86	-43.4384792722	56.0725653934	9.6878520088 8	-55.8504919068	-27.6077375522	-2.1428698976 29	-2.1428698976 29	-16.449737073	-20.450055634	-16.7060622452	-18.9347103148	suus
161	0.04	0.029414389282 5	-399.8302414 15	148.3111714 44	-38.1096483786	52.5155079562	20.4215376259	-48.7237769275	-21.7899183963	-12.2522926504	0.6731942482 94	-19.331945184 1	-17.618382395 4	-6.7214886937 4	-12.2226001813	suus
159	-0.19	0.070497690819	-317.0797452 38	115.0557226	-23.7413764763	44.2450506652	-2.9295727519 8	-34.9993817703	-1.0710389779 2	-4.3586845758 4	-19.2763278404	-13.630906986	-13.475403003 4	-23.2988585739	-0.4668861240 72	sanne
2	-0.47	0.069834351539 6	-305.6555460 56	139.2906846 86	-58.5414648373	38.7242884339	24.219689047	-37.1490617256	-21.9983621452	-14.1463703234	-20.851560206 2	-12.211334757 3	-9.3761260635 7	-25.998340113 5	-15.8062318113	job
199	-0.07	0.060212638229 1	-373.461052 34	126.8558115 91	-16.8526746544	38.5868458067	0.1024939822 3	-25.6418013696	-2.3172171720 4	-12.6637009895	-17.3071967398	-21.074024860 5	-11.340785345 1	-8.2149649013 7	-25.8615111654	suus
189	0.23	0.074138157069 7	-341.9052269 57	125.0462343 49	-14.9214633868	38.5552944625	-6.1687703611 9	-37.6438191947	-10.655041782 5	-8.471456267	-18.0017884192	-16.701915578 8	-7.3587668206 3	-11.330790478 7	-28.4019555835	suus
34	-0.31	0.041761063039 3	-364.9910139 78	139.4520191 58	-40.0546438141	46.041043258	-2.8940585249 2	-47.5233616571	-9.8396807752 4	3.4871584130 3	-13.9253431992	-22.079146926 6	-14.506944348 3	-22.6983181474	-12.5699197375	job
182	-0.22	0.062818527217	-367.437668 513	123.6540398 24	-17.2474232265	31.4041503259	2.2505327310 2	-40.1498625727	-9.9052105798 4	-7.5233372629 7	-13.8622999202	-15.151046337 9	-13.157561403 6	-13.8658440261	-18.992872003	suus
4	-0.5	0.075564205646 5	-288.4759420 87	135.6171149 905	-62.5005479829	42.3318813719	19.203069174 6	-32.1070517193	-17.2478971343	-17.8268559198	-20.017352994 1	-10.333285319 1	-9.5629490921 8	-24.1246569607	-20.2378844998	job
187	-0.37	0.065219327607 4	-341.1106246 12	124.612839 8	-23.0655599385	41.0135020382	-3.0471607421 6	-33.5820099587	-9.3716336177 6	-4.5745375515 5	-13.1494344459	-21.736726468 9	-11.367042668 6	-11.9063939848	-27.4167859934	suus
149	0.3	0.054722011935 7	-368.1792600 17	136.8975782 71	-18.9752840789	37.1292922324	-8.3834949781 3	-42.9267787659	6.6159808053 8	4.0723807152 3	-27.1914209585	-21.062419398 1	-10.815947783 3	-12.0215700847	-6.7116181530 4	sanne
147	-0.44	0.044198058545 6	-384.9242178 64	131.2982668 89	-6.6233184343 7	38.6147591544	-5.1867489598 8	-38.0342639212	6.3690602292 2	4.1663742821 5	-26.2592404204	-20.152974175 8	-13.857514135 8	-15.281645593	-8.2854304536	sanne
54	-0.04	0.059643890708 7	-354.3651848 46	132.1810222 163	-28.3020041843	54.344163447	7.4141483762 5	-38.4116724476	-26.351117369 4	-4.1790189003 7	-4.1547109230 1	-23.791613738 5	-22.008666166 8	-3.8630112742 4	-18.1676890572	karen
132	0.08	0.072326982337 2	-358.6783295 19	119.8980669 34	-18.9113514033	47.5024951633	-7.0680357832 1	-47.5024951633	7.4684211801 7	-1.9265921760 7	-22.126735552 2	-13.296159241 5	-20.260266291 9	-12.7176515681	-4.7328235211 8	sanne
77	-0.27	0.068955585360 5	-357.9362692 68	127.7650859 61	-4.4519547862 2	26.3043800387	-18.222027858 9	-37.5982011726	-4.1229140859 7	-1.2456739203	-9.8799488434 8	-29.176615410 8	-13.044460280 4	1.5157171787 59	-20.7086201003	karen

Figure 5: Screencapture of the dataset

References

Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., & O'Shaughnessy, D. (2009). Developments and directions in speech recognition and understanding, part 1 [dsp education]. *IEEE Signal Processing Magazine*, *26*(3).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chen, G., Parada, C., & Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In *Acoustics, speech and signal processing (icassp), 2014 ieee international conference on* (pp. 4087–4091).

Foote, J. (1999). An overview of audio information retrieval. *Multimedia systems*, *7*(1), 2–10.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Hasan, M. R., Jamil, M., Rahman, M. G. R. M. S., et al. (2004). Speaker identification using mel frequency cepstral coefficients. *variations*, *1*(4).

Herbig, T., Gerl, F., & Wolfgang, M. (2010). Simultaneous speech recognition and speaker identification. *Spoken Language Technology Workshop (SLT)*.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*(1), 12–40.

McFee, B., McVicar, M., Nieto, O., Balke, S., Thome, C., Liang, D., ... Lee, H. (2017). *librosa.* Retrieved from `http://librosa.github.io/librosa/`

Mitchell, T. M. (1997). *Machine learning - decision tree learning* (Vol. 45) (No. 37).

Pham, H. (2017). *Pyaudio.* Retrieved from `https://people.csail.mit.edu/hubert/pyaudio/`

Reynolds, D. A., F., Q. T., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, *10*, 19–41.

Weenink, D. (2017, 02). Personal Communication.

Weihs, C., Jannach, D., Vatolkin, I., & Rudolph, G. (2016). *Music data analysis: Foundations and applications.* CRC Press.

Zang, A. (2015). *Speech recognition.* Retrieved from `https://github.com/Uberi/speech_recognition#readme`