Spoken Language Identification in the Image Domain

Alva Liu
alvaliu@kth.se

Mikaela Åstrand
miastra@kth.se

Abstract

Spoken language identification systems (LID) allow for automatic language detection given speech data. Among the many available methods that can be applied to this classification task, modern machine learning and deep learning approaches have been reported as effective. A previous study approached the problem of spoken language identification in the image domain by transforming speech samples to spectrograms and classifying them using convolutional neural networks (CNN) (1). We have implemented two similar types of CNNs and trained them on data for five languages from the SpeechDat database. Then, we investigated how well their performance generalized on speech samples from another source then SpeechDat. The results indicated that even though the models could achieve over 80 % in test accuracy on SpeechDat data, they did not perform well on speech samples not originating from the SpeechDat database, with the best model achieving 37.5 % accuracy.

1 Introduction

The application of voice when interacting with modern technology is rapidly increasing. Many intelligent products such as Amazon Alexa, Google Translate and Apple's Siri are already applying speech recognition technologies to understand context from speech, and can subsequently be controlled by voice. A limitation to many of these products is however that they have to be explicitly told which language is being used. Spoken language identification addresses the problem of automatically determining the language being spoken given natural speech in the audio domain. Several machine learning approaches have previously been used to construct language identification systems (LID) that attempt to solve the task of correctly classifying speech samples from different languages. Some common approaches include Gaussian mixture models, support vector machines, and various types of neural networks.

1.1 Gaussian mixture model LID systems

Gaussian mixture models (GMM) operate under the assumption that different languages have different sounds and hence different sound frequencies. A GMM model is created for every language by extracting feature vector streams from training speech samples of that language and then cluster the streams, resulting in a number of cluster centers that serve as the initial estimations of the means of the Gaussian densities. Through the EM-algorithm, the model parameters are then repeatedly re-estimated until convergence. A previously unseen sample can subsequently be classified by calculating and comparing the log likelihood of it being produced by each of the language models (2).

1.2 Support vector machine LID systems

Support vector machines (SVM) have also previously been used for LID. A possible approach is to train a set of "one-against-all" linear SVMs. Each of these linear SVMs is trained to separate speech samples that belong to a specific language from the rest of the languages in the system. For an unseen speech sample, the predicted label is decided by the classifier that performs best (3).

1.3 Neural network LID systems

Neural networks and deep learning have in recent years proven to be effective at different tasks concerning speech recognition. Since speech is generally difficult to work with in the raw audio format, previous studies have instead attempted to use frequency representations of speech as input.

Gonzalez et al. proposed in 2015 (4) a real-time end-to-end multilingual speech recognition architecture based on a deep feed forward neural network as the main LID-component to classify 34 different languages. Their deep neural network was fed with mel filterbanks coefficients computed from speech data.

Bartz et al. (1) constructed a LID system in 2017 that approached the problem of language identification in the image domain by transforming speech samples into spectrograms. Their model was based on a convolutional recurrent neural network (CRNN) architecture for image-based sequence recognition originally proposed by Shi et al. (5). The CRNN architecture applies a bidirectional long short-term memory layer on top of the features extracted from the convolutional layers, thereby considering the sequential dependence in speech data. In Bartz et al.'s study, audio samples in 6 different languages from speeches, press conferences and statements from the European Parliament and news broadcast channels on YouTube were collected and used for model training. The languages were English, German, French, Spanish, Mandarin Chinese, and Russian. Their LID system achieved an overall accuracy of 92 %, performing worst on English with an accuracy of 86 % and best on Chinese with an accuracy of 96 %. Their work formed the main inspiration for this project.

1.4 Problem formulation

In this project, we constructed two types of convolutional neural networks trained on varying amounts of data from the SpeechDat database. The aim was to evaluate how well models trained on SpeechDat data solely can generalize on speech data that originate from a different source.

2 Method

Our project consisted of three parts: the transformation of multilingual speech data to the image domain, the training and evaluation of two neural network architectures for classifying SpeechDat data by language, and, finally, the recording of and evaluation of our models on new speech data.

2.1 Dataset and preprocessing

The data used in this study comes from the SpeechDat database, a database that contains speech samples from numerous European languages and variants recorded either over the fixed or the mobile network (6). We used in total $\sim$200.000 speech samples recorded over the fixed network in Swedish, English, German, Spanish and French. The last four languages were chosen to make the results comparable to those of (1).

The SoX software (7) was used to transform the audio files into grayscale spectrograms in the .png format to be used as input to the network. To have images of equal size, the audio files were cut into 5 s fragments and all shorter files were disregarded. (1) used 10 s fragments, but as described in (8), that would have excluded most of the data in at least the Swedish SpeechDat database, wherefore 5 s was chosen instead. We used the same height and resolution as them, 129 pixels and 50 pixels/s respectively, giving us images of size 129x250 pixels. Two example images can be seen in Figure 1.

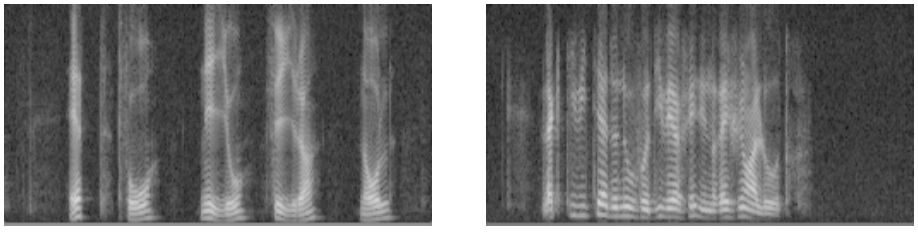


Figure 1: Two spectrogram examples.

We started with only the first CD from each language, corresponding to 200-300 speakers/language and 40-48 audio files/speaker. We then performed the previously described preprocessing on all these audio files. The resulting spectrograms were split to have approximately 70 % of the speakers of each language in the training set, 20 % in the validation set, and 10 % in the test set. Finally, each dataset was balanced to have the same number of spectrograms from each language, randomly picked from all included speakers.

We later performed the same preprocessing on the files from the second CD from each language, and merged the respective training, validation, and test set to form larger sets approximately double the size. The combined test set was used for evaluation of all the models. A summary of the number of audio files and 5 s snippets from each of the CDs can be seen in Table 1, and the number of spectrograms in each of the final datasets can be seen in Table 2. The "small" datasets include the data from CD 1 of each language, and the "large" datasets include the data from CD 1 and 2.

Table 1: Data distribution between languages.

CD	Type	Swedish	English	German	Spanish	French	Total
1	Audio files	19 200	18 400	21 374	24 000	28 621	111 595
	5 s snippets	4 064	7 326	6 043	7 426	4 739	29 598
2	Audio files	19 200	18 398	21 154	24 000	28 630	111 382
	5 s snippets	5 214	6 986	6 201	7 335	4 691	30 427

Table 2: Number of spectrograms in the balanced datasets.

Dataset	Per language	Total
Small training	2 798	13 990
Small validation	841	4 205
Large training	6 072	30 360
Large validation	1 802	9 010
Test	881	4 405

2.2 Model architecture and selection

Two deep convolutional neural network architectures were compared in this study. The first architecture consists of five convolutional layers followed by a dense layer, as shown in Table 3 (hereinafter referred to as *the CNN*). The second architecture consists of the same convolutional layers as the first architecture, followed by a bidirectional long short-term memory (BLSTM) layer, and is referred to as *the CRNN*. The CRNN architecture was originally proposed by Shi et al. in 2017 (5).

Type	Configurations
Input	250 x 129 grayscale images
Convolution	16 maps, k:7x7, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	32 maps, k:5x5, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	64 maps, k:3x3, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	128 maps, k:3x3, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	256 maps, k:3x3, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Dense	1024 units, bn
Dropout	20 %
Output	10 units

Table 3: CNN network configuration.

Type	Configurations
Input	250 x 129 grayscale images
Convolution	16 maps, k:7x7, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	32 maps, k:5x5, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	64 maps, k:3x3, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	128 maps, k:3x3, s:1x1, bn
Max Pooling	k:2x2, s:2x2
Convolution	256 maps, k:3x3, s:1x1, bn
Max Pooling	k:2x2, s:2x2
BLSTM	256 units
Output	10 units

Table 4: CRNN network configuration.

Each model setting was trained for 50 epochs with batch size 32 and the Adam-optimizer with learning rate 0.001. The model of each type giving the highest validation accuracy was chosen for the final evaluation on the test data.

2.3 Collecting additional speech data

In order to investigate the generalisability of our models on data that do not originate from SpeachDat, we evaluated them on recordings we had made ourselves. We recorded 2-3 native speakers of each language, all aged 20-30, at the same occasion for approximately 2 min each. Their task was to read a newspaper article in their language ((9), (10), (11), (12) and (13) respectively), and the recording was done in 16 kHz linear using a standard headset and the software WaveSurfer (14). The audio files were then downsampled to 8 kHz and transformed to a-law format using SoX, before spectrograms were created as described in Section 2.1. The number of speakers and spectrograms can be seen in Table 5.

Table 5: The number of recorded speakers (male/female) and resulting spectrograms for each language.

Language	**Male**	**Female**	**Spectrograms**
Swedish	1	1	49
English	1	1	49
German	0	2	49
French	3	0	72
Spanish	1	1	50

3 Results

This section presents our results and is mainly divided into two parts: the results when evaluating the models on SpeechDat test data and when evaluating them on our own recordings.

3.1 Results on SpeechDat test data

We evaluated the performances of the CNN and CRNN models trained on the small and the large data sets respectively on the same test data set. Table 6 shows the accuracy of each model. Figure 2 shows the confusion matrices and Figure 3 the precision, recall, and F1 score for all models and languages.

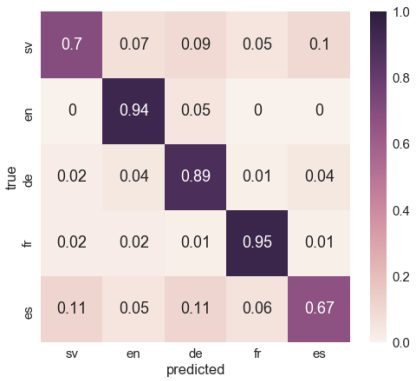
Table 6: Accuracy for the four models evaluated on the SpeechDat test data.

Model	**Accuracy [%]**
CNN trained on the small dataset	82.9
CRNN trained on the small dataset	81.6
CNN trained on the large dataset	85.2
CRNN trained on the large dataset	85.6

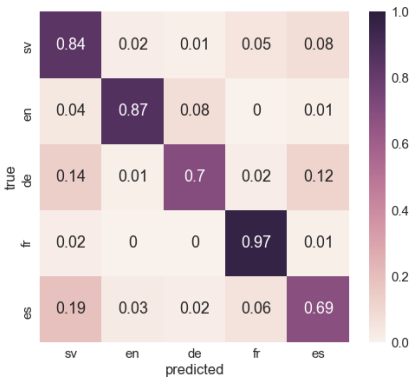
3.2 Results on our own recordings

When evaluating the models described in Section 3.1 on the spectrograms generated from our own recordings, all models classified almost all samples as German exclusively regardless of the true language. To investigate possible reasons for this, we computed the long term average spectrum (LTAS) for the different data sets. For our recordings, we concatenated all included fragments of all languages (269 in total), and for the SpeechDat data we randomly picked and concatenated 300 fragments from CD 1 of each language. The corresponding LTAS can be seen in Figure 4.

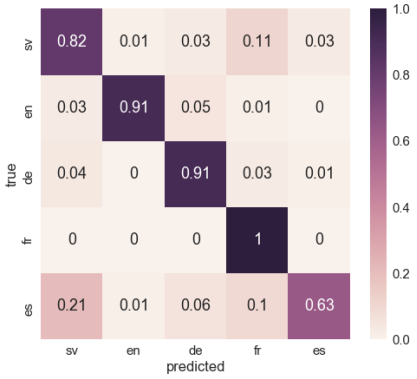
The LTAS from our recordings and SpeechDat mainly differ for low and high frequencies. We therefore decided to ignore the first and last 10 pixels in all spectrograms (corresponding to 310 Hz, marked out in Figure 4) and retrained the models on 109x250 pixels images. This yielded slightly better results for our own recordings, see Table 7, however, the models were still biased towards German. The confusion matrix for the model giving a total accuracy of 37.5 % can be seen in Figure 5, and the corresponding accuracy for each speaker can be seen in Table 8.



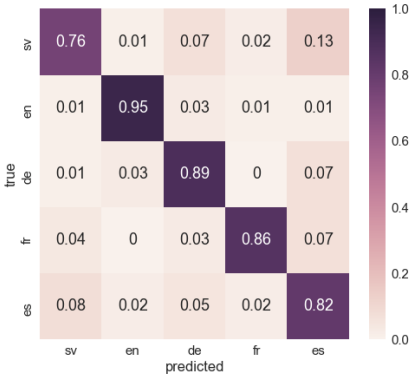
(a) CNN trained on the small dataset



(b) CRNN trained on the small dataset

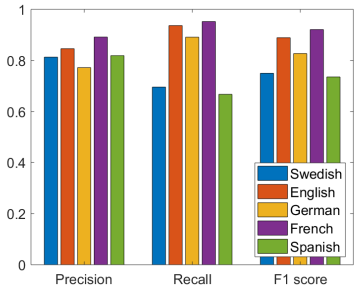


(c) CNN trained on the large dataset

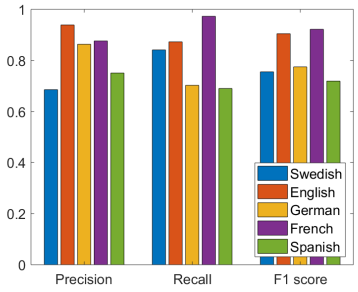


(d) CRNN trained on the large dataset

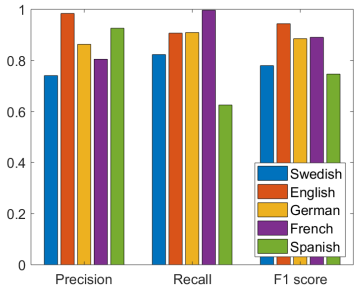
Figure 2: Confusion matrices for the four models and all languages.



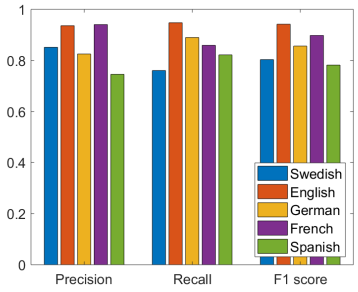
(a) CNN trained on the small dataset



(b) CRNN trained on the small dataset



(c) CNN trained on the large dataset



(d) CRNN trained on the large dataset

Figure 3: Precision, recall and F1 score for the four models and all languages.

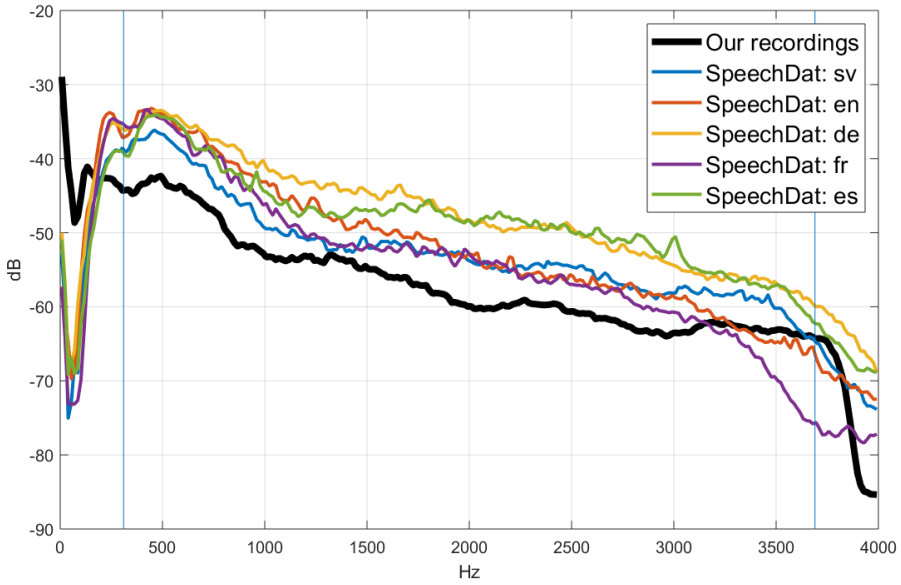


Figure 4: Long term average spectrum for all our own recordings compared to those computed on 300 5 s samples from each language from SpeechDat. The blue lines mark 310 and 3690 Hz.

Table 7: Accuracy when trained and evaluated on the cropped spectrograms.

Model	**Accuracy [%] Test Data**	**Accuracy [%] Own Recordings**
CNN trained on the small dataset	80.0	26.0
CRNN trained on the small dataset	77.1	20.4
CNN trained on the large dataset	80.2	37.5
CRNN trained on the large dataset	83.0	33.1

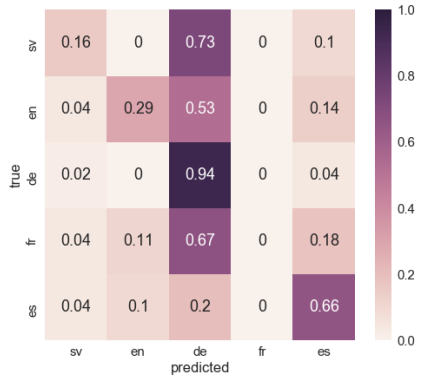


Figure 5: Confusion matrix for the model with total accuracy of 37.5 %.

Language	**Accuracy [%]**
Swedish	0
	32
English	8
	50
German	100
	88
French	0
	0
	0
Spanish	37.5
	92.3

Table 8: Accuracy for each of the 2-3 speakers of each language using the model with total accuracy of 37.5 %.

We also investigated per-image standardization by taking $(x_i - \mu_x)/\sigma_x$ of each pixel x_i, where μ_x is the mean and σ_x the standard deviation of all values in the image, for each image in the training, validation, testing set as well as our own recordings. This approach did not yield any significant changes in performance for either the test data or for our own recordings.

4 Discussion

In this section, we discuss our respective results on the different test datasets, and suggest some alternative approaches and future work building upon our project.

4.1 Evaluation on SpeechDat test data

All models performed relatively well on the test data with accuracies over 80 %, with the best model being the CNN trained on all data with an accuracy of 85.6 %. However, the performance differed between the different languages, as can be seen in Figure 2 and 3. This was also observed by Bartz et al. in their study (1), but while their models performed worst for English, it is in our case one of the better classified languages. Our model with the highest accuracy also had F1 scores between 78-94 % for all languages. Swedish had the lowest recall (76 %) due to many Swedish samples being misclassified as Spanish or German.

Both the CNN and the CRNN improved on the test data when the larger training data set was used. The CNN improved with 2.3 pp and the CRNN improved with 4 pp. Trained on the smaller dataset, the CNN performed better than the CRNN, but with more training data the CRNN outperformed the CNN albeit very slightly. This indicates that the amount of data available can be an important factor to consider when choosing between these two models, and that it might be necessary to try both of them before deciding on one. Also, the CRNN was slower to train than the CNN, but the difference was quite small so that should not have too much impact on the choice.

Even when using the larger dataset, we used only approximately 12 h of speech, while Bartz et al. (1) used two different datasets with approximately 53 h (based on EU speeches) and 540 h (from YouTube news channels) of speech. Furthermore, from the published code from their study, we discovered that they used binary-cross entropy as loss function during training instead of categorical-cross entropy. This type of error typically results in a higher test accuracy, thus we cannot use their results as a reference as they might be inadequate.

4.2 Evaluation on our own recordings

Cropping the images improved the classification performance to at least better than random, though still much worse than the performance on the SpeechDat test data. However, it differed a lot between different languages and speakers. No samples at all were classified as French, and it was still clearly biased towards German. As for the speakers, all languages but French had large differences between the different speakers. For all languages where we had one male and one female speaker (Swedish, English, and Spanish), the male speaker had the highest accuracy. The entire SpeechDat databases should be balanced by gender and include a certain proportion of speakers in each age interval (8), but we do not know if this also applies to the specific parts of the databases that we used. If not, this might explain some of the differences.

The texts that were read during our recordings were all quite different. The French article contained a lot of numbers while the others did not, and as many of the SpeechDat recordings involve series of numbers (8), we thought that in case the system learnt to recognize certain words, our French recordings would be the easiest to classify. However, this was not at all the case. Hence, it seems as if the specific words being said are not very important for the classification.

The LTAS plot in Figure 4 does not point out any clear differences between German and the other languages that could explain the bias. However, it seems as if the French average spectrum differs a lot from our recordings already from approximately 3200 Hz. This suggests that cropping the images even further to exclude more of the high frequencies could perhaps increase the number of samples being classified as French.

4.3 Alternative approaches and future work

An alternative approach that could have been taken is to classify the data at the sample level instead of at the individual spectrogram level. This would require that we keep track of which speech sample each spectrogram belongs to, and after predicting the language for all spectrograms that belong to the same sample, majority voting can be used to determine the overall prediction for the speech sample. However, since many of the samples from the SpeechDat database were between 0-10 seconds only,

the benefits of this approach are likely not substantial for SpeechDat data. For our own recordings that were each approximately 2 minutes long, this type of majority voting could have been applied but would not have yielded any good results judging from that the accuracies were so low for most speakers (see Table 8).

A suggestion on improvement for the future is to train the models with more data either from SpeechDat or from other sources that contain more variations in order to improve the models' abilities to generalize. We could also have recorded more people for a longer time per person in order to gather more data for model evaluation. Another possible improvement to the study is to explore other methods for image normalization, and investigate how they can affect the models' performances on unseen data. One approach would be some kind of normalization per speaker that could remove information related to the speakers and their recording conditions while keeping information related to the specific utterances.

Finally, instead of approaching the problem of LID in the image domain by converting the raw speech data to spectrograms, we could have attempted to process the speech data by computing the Mel filter-bank coefficients at the frame level, and used them as input to a feed-forward neural network classifier as Gonzalez et al. proposed in their study (4). For future work, one can compare the results of the two different approaches and evaluate their respective performances, advantages and disadvantages.

5 Conclusions

A relatively small amount of data was required in our experiments to achieve an overall accuracy over 80 % at classifying speech samples from the SpeechDat database in the European languages Swedish, English, German, French and Spanish. The best model was the CRNN trained on the larger dataset which performed 85.6 % accuracy on the test data from SpeechDat. However, despite performing well on SpeechDat data, neither one of the models were able to generalize on data that did not originate from the database. This was improved when some frequencies were excluded, but the best model still achieved only 37.5 % accuracy on our own recordings. The substantial discrepancy in performance between data from SpeechDat and our own recordings suggests that even though the models might be learning some underlying patterns that distinguishes the five languages, differences in background conditions are more important for the classification performance.

References

[1] C. Bartz, T. Herold, H. Yang, and C. Meinel, "Language Identification Using Deep Convolutional Recurrent Neural Networks," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 880–889.

[2] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.

[3] V. Chandrasekhar, M. E. Sargin, and D. A. Ross, "Automatic language identification in music videos with low level audio and visual features," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5724–5727.

[4] J. Gonzalez-Dominguez, D. Eustis, I. Lopez-Moreno, A. Senior, F. Beaufays, and P. J. Moreno, "A real-time end-to-end multilingual speech recognition architecture," *IEEE Journal of selected topics in signal processing*, vol. 9, no. 4, pp. 749–759, 2015.

[5] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[6] H. Höge, C. Draxler, H. v. d. Heuvel, F. T. Johansen, E. Sanders, and H. S. Tropf, "Speechdat multilingual speech databases for teleservices: across the finish line." 1999.

[7] (2015, Feb) SoX - Sound eXchange. [Online]. Available: http://sox.sourceforge.net/

[8] K. Elenius and J. Lindberg, *Documentation: FIXED1SV / FDB5000 - A 5000 Speaker Swedish Database for the Fixed Telephone Network*, Department of Speech, Music and Hearing, KTH, Mar 1999.

[9] P. Wolodarski. (2018, May 27) Rekordsiffrorna vi inte får prata om i Sverige. Dagens Nyheter. [Online]. Available: https://www.dn.se/ledare/kolumner/rekordsiffrorna-vi-inte-far-prata-om-i-sverige/

[10] A. G. Larmon. (2018, May 21) Can Art Change the World? BBC. [Online]. Available: http://www.bbc.com/culture/story/20180517-can-art-change-the-world

[11] A. Dreis. (2018, May 28) Wir helfen ihnen, sie helfen uns. Frankfurter Allgemeine Zeitung. [Online]. Available: http://www.faz.net/aktuell/sport/fussball/fussball-gibt-fluechtlingen-ein-stueck-heimat-das-beispiel-des-kreisligavereins-sv-wisper\-lorch-15603648.html

[12] F. Béguin. (2018, May 28) Forte baisse du nombre de fumeurs en France. Le Monde. [Online]. Available: https://www.lemonde.fr/sante/article/2018/05/28/forte-baisse-du-nombre-de-fumeurs-en-france_5305831_1651302.html#meter_toaster

[13] Óscar López-Fonseca and J. J. Gálvez. (2018, May 29) La Audiencia Nacional envía a prisión a Bárcenas, López Viejo y Guillermo Ortega por Gürtel. El País. [Online]. Available: https://politica.elpais.com/politica/2018/05/28/actualidad/1527512722_087578.html

[14] (2018, March) WaveSurfer. [Online]. Available: https://sourceforge.net/projects/wavesurfer/