Title: Speaker Recognition

Botond Vezse
Student

Ákos Kis
Student

Artúr Botond Balogh
Student

***Abstract*— We have chosen this project because of its versatility. Speaker recognition is a very important feature, that can be used in security, home applications and even more. If we could figure out a speaker's identity, we could tailor our services based on their personality. The goal was a model that we could use for personal applications, and we were excited about if we can create a solution based on our ideas. We wanted to find out how the length of speech affects the accuracy of the model and how many people can be handled with our system with acceptable accuracy. Our team used the VoxCeleb1 audio dataset, that is free to use and contains many people voices from YouTube videos.**

■ THE INTRODUCTION

In this project, our idea was not to use a previously made huge model with the highest accuracy, but to make our own with as little network as possible, and still be able to use it in home projects.

■ Összefoglalás

A projektet a szeleskörű felhasználhatósága miatt választottuk. A beszélő felismerés egy nagyon fontos funkció, ami a biztonságtechnikában, otthoni felhasználásban és számos más területen is alkalmazható. Amennyiben azonosítani tudnánk a beszélő személyazonosságát, személyre szabhatnánk a szolgáltatásainkat. A cél egy olyan modell megalkotása volt, ami alkalmazható lenne személyes felhasználásra, továbbá kíváncsiak voltunk, hogy létre tudunk-e hozni egy megoldást a saját ötleteink alapján, megfelelő pontosság mellett.

Tudni akartuk továbbá, hogy a beszéd hossza hogyan befolyásolja a modell pontosságát, valamint hány embert tudunk kezelni ezzel a rendszerrel.

A csapatunk a VoxCeleb1 hang adatbázist használta, ami ingyenes felhasználású és rengeteg embert tartalmaz, akik hangját YouTube videókból nyerték ki.

■ Description of the topic, previous solutions

We found various solutions on the internet that can handle huge amount of people with very high accuracy and low EER, for example the models made by Nvidia [4] managed to get 2.1% EER with cleaned VoxCeleb1 dataset (~1200 people) using SpeakerNet, and the ones made for the VoxCeleb challenge [2] (EER was between 7.8% and 10.2% with VoxCeleb2, using VGG-M) [3] (EER was 2.89% with VoxCeleb, using ECAPA-TDNN) [8] (EER was 4.59% with VoxCeleb1, using ResNet-50). We can find implementations with different data processing, where the input of the model is a time series or spectrogram

Digital Object Identifier XXXXXXXX

[1]. We chose the latter, which went along with our ideas.

■ System model

We checked multiple models according to the task it needs to achieve [5] [6] [7] then we continued with simplifying the ResNet-34 model, which was used by the Voxceleb's own solution. From our previous experience with models for speaker recognition we tried to modify the layers according to other sources while we were trying to keep the structure of model almost intact.

Model
input (129,142,1)
conv2 – 64 conv2 - 128
maxpool – k = (2,2)
dropout – 0.2
conv2 – 128 conv2 – 128
maxpool
conv2 – 64
conv2 – 64
maxpool
dropout – 0.4
flatten
FC – 1,512
dropout – 0.4
FC – number of classes
softmax

■ Implementation

We had to carefully choose the source, because we often have to deal with high noise level on the recordings, and we cannot be sure whether an audio file only contains the voice of the selected person. We found the VoxCeleb1 good enough so we could use that in our solution.

DATA PREPROCESSING

Examining the dataset, we noticed that it was very imbalanced, so we had to balance it [10]. In this task, we must process a given audio file (.wav) in a way that we could use it as input data for some kind of deep learning model. Our basic idea was to create a spectrogram from each file that contains all the information that is needed to recognize a person's voice. This information is perfectly represented in the frequency domain. We then think about the spectrograms as pictures, so the 2D convolution layers can extract information out of it. Since we had to deal with a huge dataset, we implemented data generators for this specific use case. In these generators we also dealt with the data augmentation [9] as well.

MODEL TEACHING

Teaching the models had some serious difficulty because we tried it with changing several input parameters and the results were extremely varied. The length of the audio for example made a huge difference, as it turns out the model can't be trained too much with audio that is <2s long. The number of classes made a similar impact to the system, so we ended up using 50 classes to identify.

EVALUATION

Ultimately, we hyper optimized the best model we have found and managed to make it more than 70% accurate. We ran different types of hyperparameter optimization algorithms.

TESTING

The training accuracy was 89.73%, the validation accuracy was 74.07%, and the test accuracy 73.32% (50 people, 2s audio). We managed to get 6.22% EER.

The validation accuracy, with 200 classes, went down to 0.5%, and the one with 1s of audio, to 1%.

■ Conclusion

We noticed that the accuracy is heavily depends on data preprocessing so in the future we must work on this issue. If we want to identify more people in the future, the expansion of the network will be essential. A third

way to start from is transfer learning with models which are used for this issue (for example the ResNet-34).

■ REFERENCES

1. S. Bunrit, T. Inkian, N. Kerdprasop, and K. Kerdprasop, “Text-independent speaker identification using deep learning model of convolution neural network,” Int. J. Mach. Learn. Comput., vol. 9, no. 2, pp. 143–148, 2019, doi: 10.18178/ijmlc.2019.9.2.778.
2. J. S. Chung, A. Nagrani, and A. Zisserman, “VoxceleB2: Deep speaker recognition,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2018-September, no. i, pp. 1086–1090, 2018, doi: 10.21437/Interspeech.2018-1929.
3. J. Thienpondt, B. Desplanques, and K. Demuynck, “The IDLAB VoxCeleb Speaker Recognition Challenge 2021 System Description,” 2021, [Online]. Available: http://arxiv.org/abs/2109.04070.
4 N. R. Koluguri, J. Li, V. Lavrukhin, and B. Ginsburg, “SpeakerNet: 1D Depth-wise Separable Convolutional Network for Text-Independent Speaker Recognition and Verification,” pp. 1–5, 2020, [Online]. Available: http://arxiv.org/abs/2010.12653.
5. A. S. Imran, V. Haflan, A. S. Shahrebabaki, N. Olfati, and T. K. Svendsen, “Evaluating acoustic feature maps in 2D-CNn for speaker identification,” PervasiveHealth Pervasive Comput. Technol. Healthc., vol. Part F148150, pp. 211–216, 2019, doi: 10.1145/3318299.3318386.
6. S. Yadav and A. Rai, “Learning discriminative features for speaker identification and verification,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2018-September, no. September 2018, pp. 2237–2241, 2018, doi: 10.21437/Interspeech.2018-1015.
7. I. International, W. On, M. Learning, and F. O. R. Signal, “SPEAKER IDENTIFICATION AND CLUSTERING USING CONVOLUTIONAL NEURAL NETWORKS Yanick Lukic , Carlo Vogt , Oliver D ¨ Zurich University of Applied Sciences , Winterthur , Switzerland,” 2016.
8. A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” Comput. Speech Lang., vol. 60, p. 101027, 2020, doi: 10.1016/j.csl.2019.101027.
9. E. Ma, “Data Augmentation for Audio,” 2019. https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6.
10. G. Seif, “Handling Imbalanced Datasets in Deep Learning,” 2018. https://towardsdatascience.com/handling-imbalanced-datasets-in-deep-learning-f48407a0e758.