Proceedings of the 10th International Scientific and Practical Conference «Science, Education, Innovation: Topical Issues and Modern Aspects»

(April 16-18, 2024). Tallinn, Estonia

No 196

INFORMATION AND WEB TECHNOLOGIES

Speaker identification overview

Shertayev Karim Amirovich[1]

[1] 2nd year master's student;
International University of Information Technologies; Republic of Kazakhstan

Abstract. Speaker identification is the process of finding the speaker by voice. There are two approaches to compare votes. The first is speaker identification. The second is speaker verification. Speaker verification is most often used in user authorization systems. For example, a person's voice can be a biometric method of authorization in the system. Speaker identification is used in many areas such as financial services, healthcare, hospitality and so on. From this we can conclude that the identification of the speaker is an urgent problem today. There are various ways to identify a user, ranging from conventional machine learning models to neural networks. The purpose of this article is to compare voice identification models, since choosing a suitable model is a difficult choice. Models will be compared based on many criteria, including the dataset on which they were trained.

Keywords: *Speaker identification, speaker verification, Recurrent neural network, Convolutional neural network, feature extrusion.*

Introduction. Neural networks are a powerful technology for data analysis. High processing speeds and the capacity to learn a problem's solution from a series of examples are two standout qualities of neural networks. Currently, there are two basic network models that are used in the bulk of real-world neural network applications. We provide a thorough description of these models and a breakdown of the various training methods. Neural networks solve many problems such as speech-to-text, prediction, decision making, pattern recognition, optimization, etc. This article describes the speaker identification problem.

Speaker identification is a way of finding a similar voice from a set. There is a task similar to this and this is the user verification task. The difference is that the task of identification is to find a similar voice and the task of verification is to confirm that the voice recording belongs to a particular user.

Speaker identification has become popular at the present time, as there are many applications for such technology. These areas include security, financial services, healthcare,

INFORMATION AND WEB TECHNOLOGIES

and hospitality. Neural networks with which you can identify a client by voice provide great opportunities for business, so a business can more correctly give recommendations to users and also predict many indicators.

Taking into account all these factors, it was decided to write this article. This article will be an introduction to the research topic "Speaker Identification". To begin with, let's figure out what methods and models of user identification already exist and what can be improved in them. Also an important part of becoming is extracting features from the voice recording. Data preparation is an important part of data analysis, because by extracting the right features from the data, you can improve the metrics of the model well.

Comparative analysis. The following indicators will be used to compare the works of the authors in the "Speaker Identification" task: «Dataset», «Approach», «Features», «Classification type», «Metric type», «Best score».

Dataset: The dataset used to train and test the neural networks is a crucial indicator when comparing neural networks. It is important to choose an appropriate dataset that represents the problem domain well, and to ensure that the dataset is large enough to provide a fair evaluation of the performance of the neural networks. Different datasets can have different characteristics and challenges, which can affect the performance of the neural networks.

Approach: The approach used to design and train the neural networks is another important indicator. There are various approaches to designing neural networks, such as feedforward, convolutional, recurrent, and deep learning. The choice of approach can affect the performance of the neural networks on different tasks, and can also affect factors such as training time and complexity.

Features: The features used as inputs to the neural networks are another important indicator. Features are the inputs to the neural network, which are used to make predictions or classify data. The choice of features can affect the performance of the neural network, and different types of features may be more or less suitable for different tasks.

Classification type: The type of classification task being performed is an important indicator when comparing neural networks. For example, binary classification tasks have different requirements than multiclass classification

tasks, and different types of neural networks may be better suited to different classification tasks.

Metric type: The metric used to evaluate the performance of the neural networks is an important indicator. Different metrics may be more or less appropriate for different tasks, and the choice of metric can affect the interpretation of the results. Common metrics include accuracy, precision, recall, F1-score, and area under the curve (AUC).

Best score: The best score achieved by the neural network on the evaluation metric is the final indicator. This score is used to compare the performance of different neural networks and is often used to choose the best neural network for a particular task. It is important to ensure that the best score is calculated using a fair evaluation method, such as cross-validation, to ensure that the performance of the neural network is not overestimated.

"A Deep Neural Network Model for Speaker Identification" proposes a deep neural network model for speaker identification. The authors extract Mel Frequency Cepstral Coefficients (MFCCs) and train a deep neural network with multiple hidden layers to classify speakers. The model achieved high accuracy on the task of speaker identification.

"Optical Ciphering Scheme for Cancellable Speaker Identification System" proposes an optical ciphering scheme for cancellable speaker identification. The authors use a combination of optical and digital techniques to generate a cancellable biometric template for each speaker. The proposed scheme is robust to various types of attacks and provides strong privacy protection with a speaker identification neural network.

"Open-set Speaker Identification" focuses on the problem of open-set speaker identification, where the identity of a speaker may not be known beforehand. The authors propose a novel neural network architecture called the speaker embedding network, which learns a fixed-dimensional embedding for each speaker. The model is able to identify unknown speakers with high accuracy and outperforms other state-of-the-art methods on this task.

"Speaker Recognition Based on Deep Learning: An Overview" provides an overview of recent developments in speaker recognition using deep learning. The authors review various approaches to feature extraction, network architecture, and training strategies, and compare their performance on different datasets. They also discuss open research problems and future directions in the field.

Proceedings of the 10th International Scientific and Practical Conference «Science, Education, Innovation: Topical Issues and Modern Aspects»

(April 16-18, 2024). Tallinn, Estonia

No 196

INFORMATION AND WEB TECHNOLOGIES

Table 1

Research Comparison Table

Article name	Dataset	Approach	Features	Classification type	Metric type	Best score
A Deep Neural Network Model for Speaker Identification	Aishell-1 dataset	CNN with RNN	Mel spectrogram	categorical	Accuracy = TNSV / TNTV	95.42%
Optical Ciphering Scheme for Cancellable Speaker Identification System	VoxCeleb1 dataset	CNN with RNN (ResNet34 + LSTM)	Spectrogram	Binary	False accept rate, FAR	98.35%
Open-set Speaker Identification	TIMIT dataset	UBM (Universal background model)	MFCC	Binary	metric termed Minimum Accumulative Error Rate	6.9
Speaker Recognition Based on Deep Learning: An Overview	VoxCeleb	UBM	MFCC	Binary	EER (Equal Error Rate)	4.54
Deep Learning for i-Vector Speaker and Language Recognition	NIST SRE 2006, NIST 2014	i-vector + DNN	MFCC	categorical	EER (Equal Error Rate)	6.81
Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform	-	GMM + MVN-based neural network architecture	Discrete Wavelet Transform	categorical	Correct Identification, FAR, FRR	Correct Identification = 69.1%
Speaker identification using mel frequency cepstral coefficients	Custom database	HMM	MFCC	categorical	accuracy	95.24
Support Vector Machines using GMM Supervectors for Speaker Verification	-	GMM	MFCC	categorical	accuracy	-
Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network	A 100-speaker database	HMMs, GMMs, SVMs, GMM-UBM	MFCC	categorical	accuracy	99.4

INFORMATION AND WEB TECHNOLOGIES

"Deep Learning for i-Vector Speaker and Language Recognition": In this article, the author describes the task of speaker identification and partly on language identification, since the language can influence speaker identification. Language can influence speaker identification for the following reasons.

1) Phonetic Variations: Different languages have distinct phonetic characteristics, such as sound inventories, phoneme durations, and intonations. Deep learning models for speaker identification often rely on capturing these acoustic features to discriminate between speakers. When the models are trained on one language and tested on another, the phonetic variations can introduce inconsistencies, leading to reduced performance.

2) Accent Variations: Accents within a language or across different languages can significantly impact speaker identification. Deep learning models trained on speakers with a particular accent may struggle to recognize speakers with different accents. Accents can alter the pronunciation of words, change the stress patterns, and introduce unique acoustic variations, making it challenging for the models to generalize effectively.

3) Language-Specific Acoustic Characteristics: Each language has its own set of acoustic characteristics and speech patterns. Deep learning models learn to extract relevant features from the input speech signal, such as pitch, formants, and spectral patterns, to identify speakers. These features may not generalize well across different languages, leading to reduced accuracy when the model encounters unfamiliar acoustic patterns.

4) Training Data Bias: Deep learning models require a substantial amount of labeled training data to learn representations that generalize well. If the training data is biased towards a specific language or language group, the models may not perform equally well on other languages. The lack of diversity in the training data can limit the model's ability to handle the variations introduced by different languages.

The article delves into the use of deep learning models, specifically deep neural networks (DNNs), to extract more discriminative and robust features from the input audio data for i-vector-based recognition. It discusses the advantages of deep learning in capturing complex patterns and variations in speech signals, which can potentially improve the accuracy

of speaker and language recognition systems.

"Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform":

This study discusses a hybrid approach for feature extraction from audio recordings, with a particular focus on speaker matching. The author utilizes the discrete wavelet transform (DWT) as the feature extraction technique and employs neural networks for the speaker matching process. The article highlights the advantages of this approach, emphasizing that it provides more accurate speaker characteristics compared to alternative methods.

The discrete wavelet transform is utilized to extract features from the audio recordings. This transformation divides the signal into different frequency bands, enabling the identification of localized features in both the time and frequency domains. By leveraging the DWT, the author aims to capture more discriminative and informative characteristics of the speakers, enhancing the accuracy of speaker identification.

To match speakers, neural networks are employed, taking advantage of their ability to learn complex patterns and relationships in the extracted features. Neural networks have demonstrated strong performance in various tasks, including speaker identification, due to their capacity to model nonlinear dependencies and capture intricate representations.

Furthermore, the study highlights the comparison between the wavelet transform and the discrete cosine transform (DCT), another commonly used method for feature extraction. By contrasting these two techniques, the author sheds light on the superiority of the wavelet transform in terms of accuracy and its ability to capture relevant speaker characteristics.

Speaker identification refers to the process of determining the identity of a speaker from a given audio sample. Speaker verification, on the other hand, involves authenticating an individual's claimed identity based on their voice characteristics. Both tasks rely on extracting distinctive features from the voice data and employing machine learning algorithms for classification or comparison.

The author delves into an in-depth analysis of different methods used in speaker identification and verification, including but not limited to Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and deep learning approaches like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). By

INFORMATION AND WEB TECHNOLOGIES

examining the strengths and weaknesses of these methods, the author provides valuable insights into their applicability and performance across different voice processing tasks.

Furthermore, the article presents statistics regarding the popularity of specific tasks within the realm of voice processing. This statistical analysis sheds light on the prevalence and significance of various applications, such as speaker identification in forensic investigations or voice biometrics for secure authentication systems. By understanding the landscape of voice processing tasks, researchers and practitioners can gain a better understanding of the areas where further advancements and improvements are needed.

Overall, these articles demonstrate the importance of feature extraction, network architecture, and training strategies in speaker identification. They also highlight the need for robust and privacy-preserving speaker identification systems, as well as the challenges of open-set speaker identification.

Metrics. During the comparison of research papers, it was found that the researchers used the metrics Accuracy, EER, FAR, M-AER, FRR. The most commonly used metrics are Accuracy and EER.

It is essential to consider specialized evaluation metrics designed for Speaker Identification tasks. Metrics such as Equal Error Rate (EER), Detection Error Tradeoff (DET) curve, or Rank-N Accuracy are more suitable for assessing the performance of a Speaker Identification system. These metrics take into account the trade-offs between false positives and false negatives, account for imbalanced datasets, and offer a more accurate representation of the model's capabilities in real-world scenarios. Using these metrics alongside Accuracy provides a more comprehensive and meaningful evaluation of a Speaker Identification system's performance.

How to improve the model for speaker identification? Here are a few possible suggestions for improving speaker identification by audio record:

Incorporate transfer learning: Transfer learning can be used to leverage pre-trained models from related tasks to improve speaker identification. For example, a pre-trained model for speech recognition or language modeling can be fine-tuned for speaker identification tasks.

Explore attention mechanisms: Attention mechanisms can be used to highlight important features of the audio signal,

which can be useful for speaker identification. For example, an attention mechanism can be used to focus on specific segments of the audio signal that are most informative for speaker identification.

Also I think to consider multimodal approaches: Multimodal approaches that combine audio with other modalities such as video or text can improve speaker identification. For example, lip movements or transcribed text can provide additional cues for identifying the speaker.

Overall, there is still much research to be done in the field of speaker identification, and these are just a few possible directions for my future in my work.

Conclusion. In conclusion, the articles discussed in this conversation shed light on the importance of deep learning in the development of robust and accurate speaker identification and recognition systems. These systems are crucial in various applications, including security, forensics, and voice-controlled devices.

The articles presented various techniques and approaches for speaker identification, such as deep neural network models, optical ciphering schemes, and open-set identification methods. They also discussed different metrics and evaluation methods used to assess the performance of these systems, including accuracy, EER, minDCF, FAR, and FRR.

Despite the significant progress made in speaker identification and recognition systems, challenges such as open-set identification and robustness to adversarial attacks remain. To address these challenges, it is necessary to continue exploring new techniques and evaluation methods to enhance the accuracy and robustness of these systems.

In summary, the articles provide valuable insights into the current state of speaker identification and recognition systems and emphasize the importance of continuous research and innovation in this field. With further advancements, we can expect these systems to become even more efficient and reliable, opening up new possibilities and applications in the future.

References:

[1] Furui, S., 1997. Recent Advances in Speaker Recognition, Pattern Rec. Letters. 18: 859-872.
[2] Campbell, J., 1997. Speaker Recognition: A tutorial. Proceedings of the IEEE. pp: 1437-1462.
[3] Besacier, L., J.F. Bonnastre and C. Fredouille, 2000. Localization and Selection of SpeakerSpecific Information with Statistical

INFORMATION AND WEB TECHNOLOGIES

Modeling. Speech Communications. 31: 89-106.
[4] Besacier, L. and J.F. Bonnastre, 2000. Subband Architecture for Automatic Speaker Recognition. Signal Processing. 80: 1245-1259.
[5] Damper, R.I. and J.E. Higgins, 2003. Improving Speaker Identification in Noise by Subband Processing and Decision Fusion. Pattern Recognition Letters. 24: 2167-2173.
[6] Sivakumaran, P., A.M. Ariyaeeinia and M.J. Loomes, 2003. Subband Based Text-dependent Speaker Verification. Speech Communications. 41: 485-509.
[7] Reynolds, D.A, T.F. Quatieri and R.B. Dunn., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. pp: 19-4.
[8] Bassam A. Mustafa, B. Y. Thanoon and S.D. AlShamaa., 2005. A Database System for Speaker Identificatoin. Proceedings of The 2nd International Conference on Information Technology. Al-Zaytoonah University of Jordan. May 2005.
[9] Mohd Saleem, A. M., K. Mustafa and I. Ahmad., 2005. Spoken Word of German Digits Uttered by Native and non Native Speakers. Proceedings of The 2nd International Conference on Information Technology. Al-Zaytoonah University of Jordan. May 2005.
[10] Prina Ricotti, L., 2005. Multitapring and Wavelet Variant of MFCC in Speech Recognition. IEE Proceedings on Vis. Image Signal Process., pp: 29- 35.
[11] Dokur, Z. and T. Olmz., 2003. Classification of Respiratory Sounds By using An Artificial Neural Networks. International Journal of Pattern Recognition and artificial Intelligence. 4: 567-580.
[12] Abduladheem A., M.A. Alwan, and A.A. Jassim, 2005. Hybrid Wavelet-Network Neural/FFT Nural Phoneme Recognition. Proceedings of The 2nd International Conference on Information Technology. Al-Zaytoonah University of Jordan, May 2005.
[13] Farrell, K.R., R.J. Mammone and K.T. Assalah., 1994. Speaker Recognition Using Neural. Networks and Conventional Classifiers. IEEE Trans. on Speech and Audio Proc. pp: 194-205.
[14] Ramachandran, R.P. K.R. Frrell and R.J. Mammone., 2002. Speaker Recognition- General Classifier Approaches and Data Fusion Methods . Pattern Recognition. 35: 2801-2821. 15. Burrus, C., R. Gopinath and H. Guo. 1998. Introduction to wavelets and wavelet Transforms.1st edition. Prentice Hall.
[15] Douglas A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.