Universidad Autónoma de San Luis Potosí
Facultad de Ingeniería
Centro de Investigación y Estudios de Posgrado

Artificial Neural Networks for Speaker Verification

Que para obtener el grado de:
Maestría en Ingeniería de la Computación

Presenta:
José Luis Medellín Garibay

Asesor:
Juan Carlos Cuevas Tello

San Luis Potosí, S. L. P. Enero de 2025

Abstract

Speech recognition systems are becoming popular because nowadays is easy to capture audio signal through mobile devices. Moreover, biometric recognition systems are also popular because fingerprints, voice, retina and face are specific to individuals and cannot be stolen easily as passwords and PINs. There are many dataset for speech recognition tasks. This research focus on the benchmark for speaker verification SRE-08. There are different approaches for the automatic speaker verification task: prediction, regression and classification. This research focus only on classification with Probabilistic Neural Networks (PNN) and Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Network (ECAPA-TDNN) methods. We use the Equal Error Rate (ERR) metric to measure the performance, and we found that the best performance corresponds to the classification approach (PNN) with the average EER = 13.3% (Accuracy 88%). The classification approach (PNN) achieves the single best with EER=0% (Accuracy 100%) on five speakers out of nine.

Resumen

Los sistemas de reconocimiento de voz están ganando popularidad porque, hoy en día, es fácil capturar señales de audio a través de dispositivos móviles. Además, los sistemas de reconocimiento biométrico también son populares porque características como las huellas dactilares, la voz, la retina y el rostro son específicas de cada individuo y no pueden ser robadas tan fácilmente como las contraseñas y los NIP. Existen muchos conjuntos de datos para tareas de reconocimiento de voz. Esta investigación se centra en el conjunto de datos para la verificación de hablantes SRE-08. Hay diferentes enfoques para la tarea de verificación automática de hablantes: predicción, regresión y clasificación. Esta investigación se enfoca únicamente en la clasificación utilizando Redes Neuronales Probabilísticas (PNN) y el método de Atención, Propagación y Agregación de Canales Destacados en Redes Neuronales de Retardo Temporal (ECAPA-TDNN). Usamos la métrica de Tasa de Error Igual (EER, por sus siglas en inglés) para medir el rendimiento, y encontramos que el mejor desempeño corresponde al enfoque de clasificación (PNN) con un EER promedio de 13.3.

UASLP
Universidad Autónoma de San Luis Potosí

FACULTAD DE INGENIERÍA

21 de noviembre de 2024

DR. JUAN CARLOS CUEVAS TELLO
P R E S E N T E.

Por medio de la presente me permito informarle, que en sesión ordinaria del H. Consejo Técnico Consultivo celebrada el día 21 de noviembre del presente, fue analizada su petición en la cual solicitó autorización para que el **Ing. Jose Luis Medellin Garibay** de la **Maestría en Ingeniería de la Computación**, se titule mediante la modalidad: **Publicación de Artículo en Congreso Internacional con Arbitraje o en Revista Indizada**, con el articulo denominado: **"Redes Neuronales Artificiales para la Verificación de Personal por Voz"** (Artificial Neural Networks for Speaker Verification)**.**

Al respecto, me permito informarle que su solicitud fue aprobada de conformidad, de acuerdo a las evidencias que avalan el requerimiento de titulación antes mencionada.

Sin otro particular de momento, le reitero las seguridades de mi atenta y distinguida consideración.

"MODOS ET CUNCTARUM RERUM MENSURAS AUDEBO"

A T E N T A M E N T E

UNIVERSIDAD AUTONOMA
DE SAN LUIS POTOSI
FACULTAD DE INGENIERIA
SECRETARIA

DR. RICARDO ROMERO MÉNDEZ
SECRETARIO DEL CONSEJO.

www.uaslp.mx

Av. Manuel Nava 8
Zona Universitaria • CP 78290
San Luis Potosí, S.L.P.
tel. (444) 826 2330 al39
fax (444) 826 2336

Copia. H. Consejo Técnico Consultivo.
*etn.

Contents

1 Paper

This chapter includes the paper presented at the prestigious 8th World Conference on Smart Trends in Systems, Security, and Sustainability (WorldS4 2024), held in London, United Kingdom. The conference serves as a global platform for researchers, practitioners, and academics to exchange innovative ideas, share cutting-edge research findings, and discuss the latest advancements in the fields of systems, security, and sustainability.

The presented paper highlights significant contributions to these domains, showcasing advancements that address critical challenges and propose innovative solutions. The content of this chapter encapsulates the core ideas and methodologies discussed during the conference presentation, providing readers with a detailed understanding of the work.

Additionally, a certificate of participation issued by the conference organizers has been included as a testament to the formal acceptance and successful presentation of the research. This certificate serves as recognition of the paper's quality and its contribution to the global academic and professional community.

Through its inclusion in this chapter, the research aims to inspire further exploration, foster collaboration among professionals in related disciplines, and contribute to the ongoing dialogue around systems, security, and sustainability.

CERTIFICATE

THIS IS TO CERTIFY THAT

JOSE LUIS MEDELLIN-GARIBAY

HAS DIGITALLY PRESENTED A PAPER TITLED

ARTIFICIAL NEURAL NETWORKS FOR SPEAKER VERIFICATION

AT THE

8th WORLD CONFERENCE ON

SMART TRENDS IN SYSTEMS, SECURITY AND SUSTAINABILITY (WORLDS4 2024)

LONDON, UK

SIMON JAMES FONG
CONFERENCE CHAIR

8TH EDITION WORLDS4 2024

DHARM SINGH JAT
TPC CHAIR

NILANJAN DEY
TPC CHAIR

23 - 26 JULY, 2024
DIGITAL PLATFORM : ZOOM

AMIT JOSHI
ORGANIZING SECRETARY

Organised & Managed By

GR FOUNDATION

GR SCHOLASTIC

Publication Partner

Springer

SPRINGER NATURE

Artificial Neural Networks for Speaker Verification

Jose Luis Medellin-Garibay[1,2] and Juan C. Cuevas-Tello[1]

[1] Universidad Autonoma de San Luis Potosi, Facultad de Ingeniería, Av. Dr. Manuel Nava No. 8, Zona Universitaria, San Luis Potosi, SLP, CP 78290 México
[2] Vitruvi Software, 734 7 Ave SW 3rd Floor, Calgary, AB, T2P3P8 Canada;
luis.medellin@vitruvisoftware.com, cuevas@uaslp.mx

Abstract. Speech recognition systems are becoming popular because nowadays is easy to capture audio signal through mobile devices. Moreover, biometric recognition systems are also popular because fingerprints, voice, retina and face are specific to individuals and cannot be stolen easily as passwords and PINs. There are many dataset for speech recognition tasks. This research focus on the benchmark for speaker verification SRE-08. There are different approaches for the automatic speaker verification task: prediction, regression and classification. This research focus only on classification with Probabilistic Neural Networks (PNN) and Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Network (ECAPA-TDNN) methods. We use the Equal Error Rate (ERR) metric to measure the performance, and we found that the best performance corresponds to the classification approach (PNN) with the average EER = 13.3% (Accuracy 88%). The classification approach (PNN) achieves the single best with EER=0% (Accuracy 100%) on five speakers out of nine.

Keywords: Speaker Verification, Speaker Recognition, Artificial Neural Networks, PNN, ECAPA-TDNN, Mel Frequency Cepstral Coefficients

1 Introduction

The main motivation of this research are the biometric recognition systems [9]; biometric cues such as fingerprints, voice and face are specific to an individual. Contrary to passwords and PINs, the use of biometric recognition systems have many advantages including that cannot be forgotten and cannot easily stolen. The human speech is the least obtrusive biometric measure, and it is simple to acquire thanks to the pervasive wave in society [18]. In recent years the smart devices technology has been increased extensively, in fact there are 294 million of smartphone users in 2020 [1] and 37 percent of households in the United States owned a smart home device in 2020 [2]. Voice-controlled smart devices require

[1] https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/
[2] https://www.statista.com/topics/6201/smart-home-in-the-united-states/

improved user authentication and security due to their internet connectivity and control over smart home and other devices.

Speech processing has various applications including speech recognition, language identification, and speaker recognition. Sometimes, additional information is stored and associated to speech. Therefore, Speaker Recognition can be either text dependent or text independent. Moreover, Speaker Recognition involves different tasks such as [3]: i) speaker identification, ii) speaker detection, iii) speaker diarization, and iv) *Speaker Verification* (SV).

This research focus on the latter. Automatic Speaker Verification (ASV) is defined as the use of a machine to verify a person's claimed identity from his voice [3]. Identity claims involve using an employee number, smart card, and other methods. ASV is crucial in biometrics for access-control applications, differing from speaker identification, which determines if the speaker is a specific person or part of a group.

The benchmark for *Speaker Verification* is ruled by the Linguistic Data Consortium[3] (LDC) and the National Institute for Standards and Technology (NIST). It is called the NIST Speaker Recognition Evaluation[4] (SRE), which focus on text independent speaker recognition tasks. So the LDC collected multiple telephone calls, which were audited for language, speaker identity and other features, see Cieri et al. [5]. Other important speech recognition benchmarks include YOHO [14] and TIMIT[5] [10]. More datasets for speech recognition can be consulted in C. Ceri and M. Liberman [6].

There are different types of feature extractors across the literature, and this research focus only on the widely used Mel-Frequency spaced Cepstral Coefficients (MFCCs) [7, 18]. Nevertheless, some research includes *i*-vectors as feature extractors and perceptual linear prediction (PLP) features [2]. The state-of-the-art algorithms for ASV includes Gaussian Mixture Models (GMM) [18, 2]. Tables 1 summarizes papers with few speakers comparing accuracy, methods and feature extraction techniques. Cloud services from Google, IBM, Microsoft, and other providers are promising but can be complex to setup and often lack publicly disclosed accuracy metrics.

This paper cover Artificial Neural Networks (ANN): Probabilistic Neural Networks (PNN) [17] and ECAPA-TDNN [8], which are classification methods. The main contribution of this paper is the methodology based on PNN for speaker verification, and also the comparison between PNN and ECAPA-TDNN on SRE data. The results show that the best results stand on PNN (classification).

This paper is organized as follows: the audio data (training and testing) at §2, speech features at §3, proposed approaches and methods are in §4, results and conclusions at the end.

[3] https://www.ldc.upenn.edu/
[4] https://www.nist.gov/itl/iad/mig/speaker-recognition
[5] https://catalog.ldc.upenn.edu/ldc93s1

Table 1. Review on publications with datasets studying few speakers. Comparing the Accuracy (Acc).

Method	Metric (Acc)	Feature extraction	Speakers - Dataset	Ref.
Vector Quantization, k-means, Euclidean distance	88.8%	MFCC	10 speakers, 5 samples for training, own dataset one sample for testing	[12]
Backpropagation (BPNN)	92%	MFCC	10, 20, 30, 40 and 50 users, own dataset	[20] *
SCNN (Siamese Convolutional Neural Network): VGG and ResNet 50	81.2%	Spectogram with FFT	10 speakers, from LibriSpeech database	[19]
DNN	43.6% - 58.9%	T-MFCC, i-vectors	7 speakers, own dataset	[1]
Backpropagation	93.1%	MFCC	14 speakers, own dataset	[11]
PNN ECAPA-DNN	41% - 88%	MFCC	9 speakers for training, and 111 for testing (SRE-08)	**This**

* text-dependent

2 Audio Data

As we mentioned above, the audio recordings are obtained from NIST-SRE[6]. The audio samples are represented as time-domain waveforms, and they correspond to the NIST Year 2008 Speaker Recognition Evaluation Plan[7] (SRE-08). The audio samples have a duration of 2 minutes on average.

NIST-SRE provides audio samples for training, and it also provides a different set of audio samples for testing. Training samples contain only target speakers, and test samples have both target and non-target (impostor) speakers. So researchers can measure the performance of their speaker recognition model.

2.1 SRE-08

The SRE-08 is limited to the **speaker detection** task, which is to determine whether a specific speaker is speaking during a given segment of conversational speech [13]. The speaker detection task is divided into thirteen distinct and separate tests involving six training conditions and one or four test conditions. There are 1,270 speakers for training and 98,777 for testing. This research focus only on the core test, i.e. the following combination: the training condition *short2* and the test condition *short3*. Both conditions consist of a two-channel telephone

[6] https://www.nist.gov/itl/iad/mig/speaker-recognition
[7] https://catalog.ldc.upenn.edu/LDC2011S08

conversational excerpt of approximately five minutes total duration. The conversation involves the target speaker and an interviewer (non-target) [13]. The datasets are given by NIST in SPHERE[8] audio files. NIST and LDC have tools for converting SPHERE audio files in conventional formats such as WAV files[9].

2.2 Dataset

In Table 2 we show the training list[10] corresponding to the training condition *short2* and the test condition *short3* of SRE-08 [13]. Additionally, it is not expected to examine performance results over all trials of the core test[11]. Therefore, common conditions can be considered in order to build a subset of the core test trial. We consider the following condition "All trials involving only English language telephone speech spoken by a native U.S. English speaker in training and test" [13]. Moreover, we only evaluate male voice ('Sex = m', in Table 2). Therefore, our training data contain speech from nine speakers.

Table 2. SRE-08 Training Data and Test Cases

Model Id	Sex	Segment Id: Channel	Speaker Id	Speech type	Channel type	Lang.	Speaker native language	#test cases	#targets
10371	m	tljiv:a	105088	phonecall	phn-main	ENG	USE	11	2
12459	m	ttqco:a	110372	phonecall	phn-main	ENG	USE	10	1
12499	m	trbfl:a	110667	phonecall	phn-main	ENG	USE	19	1
13076	m	tahak:b	103394	phonecall	phn-main	ENG	USE	14	4
13845	m	tbjem:a	111590	phonecall	phn-main	ENG	USE	15	1
14692	m	txkzs:b	103675	phonecall	phn-main	ENG	USE	10	4
14856	m	txqmx:b	102578	phonecall	phn-main	ENG	USE	14	5
15604	m	ttwkz:b	107851	phonecall	phn-main	ENG	USE	9	1
16295	m	tlskl:b	104179	phonecall	phn-main	ENG	USE	9	2
							Total	111	21

Table 2 shows the entire dataset employed in this research. There are nine speakers for training, and there are 111 test speakers, voice excerpts, with only 21 targets. For example, the speaker one, Model Id 10371, has 11 test cases with 2 targets, where 9 are impostors.

3 Speech Features

The first step for most speech recognition systems is feature extraction from the time-domain sampled acoustic waveform (audio); see Fig. 1a. The time-domain

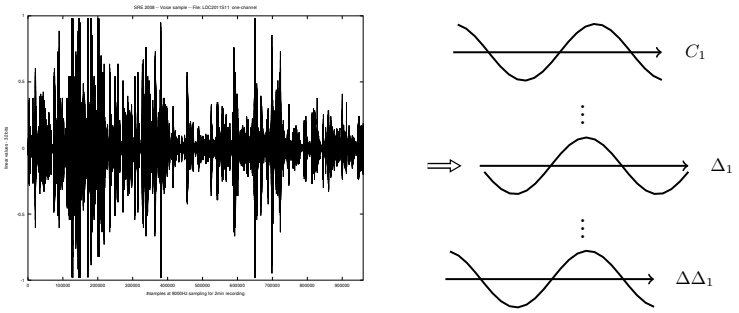
[8] NIST provies software for manipulating SPHERE files; see http://www.nist.gov/itl/iad/mig/tools.cfm

[9] https://www.ldc.upenn.edu/language-resources/tools/sphere-conversion-tools

[10] The full training list is provided by NIST through the following file NIST_SRE08_short2.model.key

[11] The list of testing data is provided by the file NIST_SRE08_short2-short3.trial.key

waveform is divided into overlapping *frames*, each generated every 10ms with a duration of 25ms. A feature is then extracted for each frame. Several methods for feature extraction (acoustic representation) have been explored, including Linear Prediction Coefficients (LPCs), Perceptual Linear Prediction (PLP) coefficients, and Mel-Frequency Cepstral Coefficients (MFCCs). The latter is widely used and some researches proofed that are the best option for feature extraction [7]. MFCCs are based on the Fast Fourier Transform (FFT), Mel-spaced filter bank values and Cepstral analysis [18]. In order to capture dynamic information between frames, the MFCCs vector is concatenated with the first (Δ), second ($\Delta\Delta$) and even third order derivative ($\Delta\Delta\Delta$) approximations [7, 18]. There are several tools for generating MFCCs including Voicebox[12] and HTK MFCC MATLAB® by Kamil Wojcicki[13].



(a) Audio (b) MFCCs

Fig. 1. Audio sample. (a) Audio represented as a time-domain waveform; (b) A set of MFCCs is obtained from the time-domain waveform.

The speech features used along this paper correspond to MFCCs; see Figure 1b. We denote standard MFCCs as $C_1, C_2, \cdots, C_N$; first order derivative MFCCs as $\Delta_1, \Delta_2, \cdots, \Delta_N$; and second order derivative MFCCs as $\Delta\Delta_1, \Delta\Delta_2, \cdots, \Delta\Delta_N$; where N is the number of coefficients. Therefore, a single time-domain waveform is represented by

$$\mathbf{T} = \begin{pmatrix} C_1^1 & C_1^2 & \cdots & C_1^M \\ C_2^1 & C_2^2 & \cdots & C_2^M \\ \vdots & \vdots & \cdots & \vdots \\ C_N^1 & C_N^2 & \cdots & C_N^M \\ \hline \Delta_1^1 & \Delta_1^2 & \cdots & \Delta_1^M \\ \Delta_2^1 & \Delta_2^2 & \cdots & \Delta_2^M \\ \vdots & \vdots & \cdots & \vdots \\ \Delta_N^1 & \Delta_N^2 & \cdots & \Delta_N^M \\ \hline \Delta\Delta_1^1 & \Delta\Delta_1^2 & \cdots & \Delta\Delta_1^M \\ \Delta\Delta_2^1 & \Delta\Delta_2^2 & \cdots & \Delta\Delta_2^M \\ \vdots & \vdots & \cdots & \vdots \\ \Delta\Delta_N^1 & \Delta\Delta_N^2 & \cdots & \Delta\Delta_N^M \end{pmatrix} \tag{1}$$

[12] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[13] http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab

where M is the number of *frames* obtained from the time-domain waveform. Each column in $\mathbf{T}$ represents the total number of MFCCs, and each row contains the total number of frames per coefficient.

4 Approaches

There are several approaches for speech recognition including prediction, regression and classification. This research only covers classification for the text independent speaker recognition task [13].

4.1 Classification

For classification, the data is represented by $\mathbf{P}$ and $\mathbf{Tc}$, see Eq. (2).

$$D_1 \begin{pmatrix} \mathbf{P} = & \underset{[(N\times 3)\times M]}{\mathbf{T_1}} & \underset{[(N\times 3)\times M]}{\mathbf{T_2}} & \cdots & \underset{(N\times 3)\times M}{\mathbf{T_U}} \\ & & & \vdots & \\ & \Downarrow & \Downarrow & \vdots & \Downarrow \\ \mathbf{Tc} = & \underset{[1\times M]}{class=1} & \underset{[1\times M]}{class=2} & \cdots & \underset{[1\times M]}{class=2} \end{pmatrix} \tag{2}$$

Each matrix $\mathbf{T_1}, \mathbf{T_2}, \cdots \mathbf{T_i} \cdots \mathbf{T_U}$ contains the MFCCs of a single user, see Eq. (1), where U is the number of training users (speakers). The dimension of $\mathbf{T_i}$ is $[(N \times 3) \times M]$. $\mathbf{Tc}$ contains the class for each user $\mathbf{T_i}$ (the labels). For illustration purposes, Eq. (2) onlys shows the data (D_1) where the target user is assigned to $class = 1$, and all non-target users to $class = 2$. The full data for classification is represented by $\mathbf{D} = D_1, D_2, \cdots D_i, \cdots D_U$.

4.2 PNN

The first classification method is Probabilistic Neural Network (PNN). The PNN architecture is shown in [17]. The input layer have $\vec{X} = x_1, x_2, \cdots, x_i, \cdots x_q$ inputs. At the pattern layer, the functions ϕ_i represent the basis functions; as in Radial Basis Function (RBF) neural networks $\phi_i = exp\left(\frac{-\|\vec{X}-x_i\|^2}{2\sigma^2}\right)$, where $\|\cdot\|$ denotes the Euclidian distance and σ the spread, which is the only free parameter. Therefore, $f_A(\vec{X}) = S_w/S_s$ and $f_B(\vec{X}) = S_w/S_s$, which is a normalized linear combination of Gaussian functions. Here f_A and f_B represent only two categories (classes) for illustration purposes, so if there are only two classes then the outputs are binary. The output o_i is estimated with f_A and f_B, and there is an algorithm that determines if the input $\vec{X}$ belongs to the category A, B or none [17]. However, the PNN can have $o_1, o_2, \cdots, o_i, \cdots, o_k$ classes (labels). The more classes, the more hidden units.

4.3 ECAPA-TDNN

The second classification method is Emphasized Channel Attention, Propagation, and Aggregation in TDNN (ECAPA-TDNN) [8]. This architecture is designed to enhance the performance of speaker verification systems that rely on

Time-Delay Neural Networks (TDNNs). It is based on the original x-vector architecture [16], but ECAPA-TDNN put more Emphasis on Channel Attention. ECAPA-TDNN incorporates several key components to achieve its goals:

- Channel Attention Mechanism: This mechanism emphasizes relevant information extracted from various acoustic channels, allowing the model to concentrate on important features and acoustic cues. This enhances the model's ability to distinguish between speakers.
- Information Propagation: The architecture incorporates mechanisms for efficient information propagation, ensuring that relevant information is accurately transmitted and processed throughout the network. This enables the model to capture essential speaker-specific characteristics.
- Aggregation Methods: ECAPA-TDNN employs advanced aggregation techniques to combine information from different parts of the network. This enhances the system's overall performance and robustness, making it well-suited for challenging speaker verification tasks.

5 Experiments and Results

Two approaches were introduced in §4.1, the best performance is achieved by the approach where all target users are assigned to $class = 1$, and all non-target users to $class = 2$ (impostors). Therefore, this approach is used for PNN and ECAPA-TDNN classification methods.

5.1 PNN

The spread parameter is set to $\sigma = 0.1$ for training and testing. There is a PNN per speaker, and the input and output data is as described in Eq. 2; where $N = 16$ and $M = 3000$ frames. We performed our experiments with MATLAB®, function `newpnn`.

Once the PNN is built and trained per speaker, then the testing part is performed by feeding the frames per each test case. Since we have a PNN per speaker, the PNN with the maximum number of frames classified as $class = 1$ is selected. In order to be consistent with previous approaches, the number of frames classified as target is represented with a number between $M_{class} = [0, 1]$ and then converted as a minimization problem. Thus, $error = 1 - M_{class}$. The results using EER per speaker are in Table 3, the average EER is 13.3%.

5.2 ECAPA-TDNN

The methodology employed for classification in this study took a distinctive departure from conventional approaches. In contrast to PNN, and similar to x-vectors, there was no requirement for audio frame splitting. Instead, we directly processed the audio input utilizing the powerful ECAPA-TDNN architecture, emphasizing the extraction of discriminative features through time delay neural networks (TDNN) with contextual awareness.

Table 3. EER: Results from PNN and ECAPA-TDNN

Speaker	Model	PNN			ECAPA-TDNN		
	Id	EER (%)	Thresh.	Acc.	EER (%)	Thresh.	Acc.
1	10371	0.0	0.78	1.00	34.31	0.20	0.70
2	12459	33.3	0.82	0.82	41.24	0.06	0.57
3	12499	22.2	0.78	0.78	0.00	0.23	0.99
4	13076	0.0	0.74	1.00	40.59	0.19	0.55
5	13845	35.7	0.84	0.66	34.03	0.02	0.68
6	14692	0.0	0.70	1.00	58.13	0.17	0.48
7	14856	0.0	0.78	1.00	40.00	0.22	0.55
8	15604	0.0	0.77	1.00	51.67	0.16	0.53
9	16295	28.5	0.82	0.77	69.64	0.27	0.39
	Average	13.3		0.88	41.39		0.60

A noteworthy aspect of our approach is the meticulous fine-tuning of parameters, with special attention given to those inherent to ECAPA-TDNN, thereby enhancing the model's discriminative capabilities. For this research endeavor, we made a deliberate choice to harness the default configurations of ECAPA-TDNN available within the Speechbrain toolkit, which is a comprehensive open-source platform dedicated to advanced speech and audio processing [15].

It is important to highlight that our decision to align with the default ECAPA-TDNN configurations in SpeechBrain not only underscores our commitment to methodological transparency but also aligns with contemporary trends in the field. The implementation was conducted using Python 3, leveraging the robust capabilities of the SpeechBrain library, which provides extensive support for cutting-edge speech and audio processing tasks. The results are in Table 3.

6 Discussion

The use of well studied benchmarks such as SRE-08 gives the advantage that the dataset is already divided into training and testing, see Table 2. Thus, the testing part is accurate because this audio was not used during training. From the point of view of machine learning, knowing the ground truth is possible to measure bias and variance and other performance metrics.

ECAPA-TDNN have reported outstanding performance in literature with EER=0.87%-1.22% [8]. Nevertheless, this performance is achieved using VoxCeleb dataset [4]. Comparing the quality of the audio between SRE-08 and VoxCeleb, the SRE-08 is more challenging because VoxCeleb does not have background noise and other real conditions when recording audio.

This research compares PNN with ECAPA-DNN on the same dataset, SRE-08, see Table 2. The best performance achieved so far is for the classification approach and the PNN method with the average EER = 13.3% (Accuracy 88%), see Table 3. These results are comparable to those reported by other researchers using similar datasets, see Table 1.

7 Conclusions

Finally, we believe the best performance is achieved by PNN, with an average EER of 13.3% (88% accuracy). This is because a model for each speaker is built and trained considering both the target speaker and non-targets (impostors), allowing the model to learn which frames belong to the target and which belong to impostors. Moreover, PNN achieved a perfect EER of 0% (100% accuracy) for five out of nine speakers. In contrast, ECAPA-TDNN achieved a perfect EER of 0% for only one speaker. Therefore, PNN demonstrates superior performance with this dataset with few speakers. However, we anticipate that ECAPA-TDNN will achieve better results with the full SRE-08 dataset, as deep learning approaches typically perform better with larger datasets.

8 Future work

Further research can explore several directions. One is to use the full SRE-08 dataset for all evaluation conditions, encompassing 1,270 speakers and 98,777 testing cases. Given that PNN is computationally demanding for testing, parallel computing will be required. Additionally, it is worthwhile to investigate the optimal number of frames per user to minimize training and testing time. Using fewer frames will expedite these processes. Moreover, the PNN's spread parameter (σ) was fixed for all speakers; setting this parameter individually for each speaker could potentially improve results. Since deep learning methods improve with larger datasets, comparing ECAPA-TDNN with PNN using the full SRE-08 dataset is a promising area for further study. Additionally, this research only utilized MFCCs as the feature extraction technique. Future work could explore other techniques such as LPCs, PLP, and i-vectors, as well as hybrid models combining PNN and ECAPA-TDNN, using the proposed approaches on the SRE-08 dataset. Deploying in real-time applications, like smart devices, poses significant computational challenges. Cloud computing could provide a substantial advantage in overcoming these limitations.

Acknowledgments We extend our gratitude to Dr. Juan Nolazco for introducing us to this challenging research area and for suggesting this new avenue in speech recognition related to x-vectors.

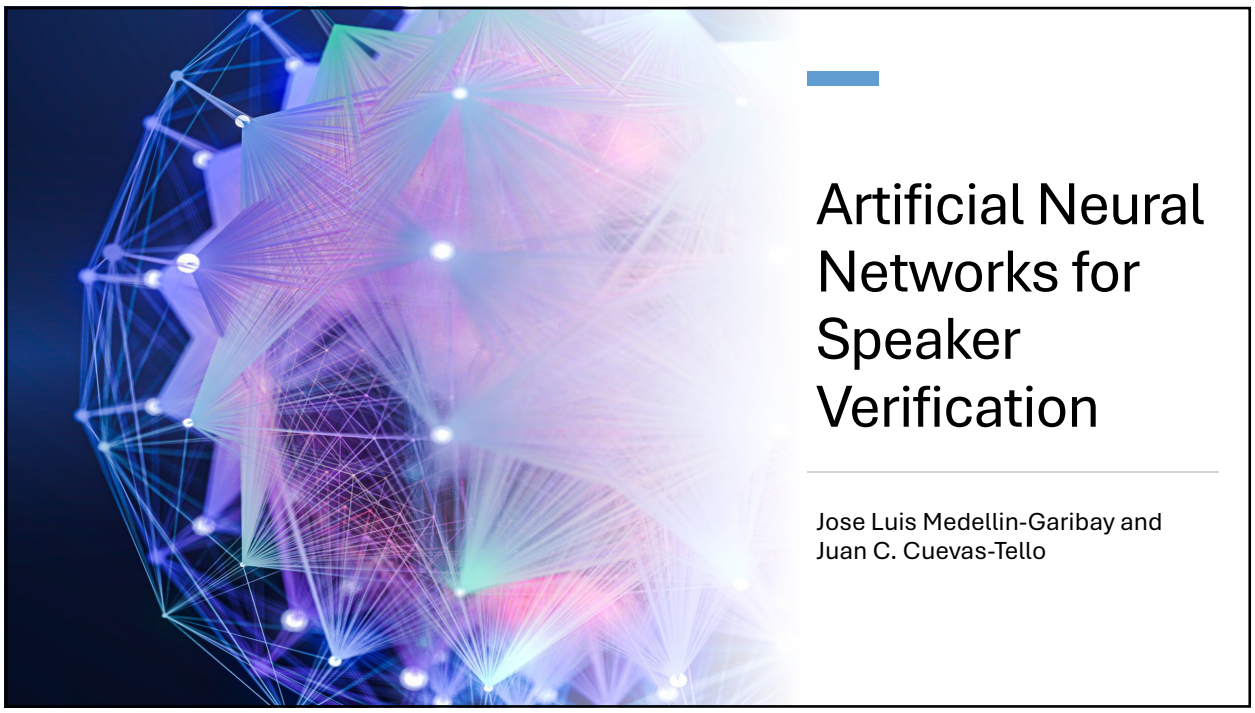
References

1. Aizat, K., Mohamed, O., Orken, M., Ainur, A., Zhumazhanov, B.: Identification and authentication of user voice using dnn features and i-vector. Cogent Engineering **7**(1), 1751,557 (2020). DOI 10.1080/23311916.2020.1751557
2. Alam, M., Kinnunen, T., Kenny, P., Ouellet P.and O´Shaughnessy, D.: Multitaper MFCC and PLP features for speaker verification using i-vectors. Speech Communication **55**(1), 237–250 (2013)

3. Campbell, J.P.: Speaker recognition: a tutorial. Proceedings of the IEEE **85**(9), 1437–1462 (1997). DOI 10.1109/5.628714
4. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
5. Cieri, C., Liberman, M.: Data for empirical foundations of forensic linguistics. In: Proceedings of LSA Symposium (2014)
6. Cieri, C., Liberman, M.: Dimensions of Speaker Recognition Research Data. LSA Symposium: Data for Empirical Foundations of Forensic Linguistics (2014)
7. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28 (1980). DOI 10.1109/TASSP.1980.1163420
8. Desplanques, B., Thienpondt, J., Demuynck, K.: Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143 (2020)
9. Fazel, A., Chakrabartty, S.: An overview of statistical pattern recognition techniques for speaker verification. IEEE Circuits and Systems Magazine **11**(2), 62–81 (2011). DOI 10.1109/MCAS.2011.941080
10. Hinton, G., Deng, L., Yu, D., Dahl, G.E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine **29**(6), 82–97 (2012). DOI 10.1109/MSP.2012.2205597
11. Kaphungkui, N., Kandali, A.: Classification and real time testing of speaker recognition system. International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET) (6) (2020)
12. Manjula, G., M., S.K., M, B.: Speaker recognition using cepstral analysis. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH AND TECHNOLOGY (IJERT) **4**(8) (2015). DOI 10.17577/IJERTV4IS080161
13. NIST: The NIST Year 2008 Speaker Recognition Evaluation Plan (2008). URL http://www.itl.nist.gov/iad/mig/tests/sre/2008/
14. Ozaydin, S.: Design of a text independent speaker recognition system. In: 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), pp. 1–5 (2017). DOI 10.1109/ICECTA.2017.8251942
15. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al.: Speechbrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624 (2021)
16. Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., Khudanpur, S.: Speaker recognition for multi-speaker conversations using x-vectors. In: ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp. 5796–5800. IEEE (2019)
17. Specht, D.: Probabilistic Neural Networks. Neural Networks **3**(1), 109–118 (1990)
18. Togneri, R., Pullella, D.: An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine **2**(1), 23–61 (2011)
19. Vélez, I., Rascón, C., Pineda, G.F.: One-shot speaker identification for a service robot using a cnn-based generic verifier. ArXiv **abs/1809.04115** (2018)
20. Wali, S.S., Hatture, S.M., Nandyal, S.: Mfcc based text-dependent speaker identification using bpnn. International Journal of Signal Processing Systems **3**(1), 30–34 (2015). DOI 10.12720/ijsps.3.1.30-34

2 Presentation

This section includes the presentation slides associated with the paper presented at WorldS4 2024. The presentation highlights the key aspects and findings of the research in a concise format designed for conference attendees.



Artificial Neural Networks for Speaker Verification

Jose Luis Medellin-Garibay and
Juan C. Cuevas-Tello

1

Introduction

Motivation: Biometric recognition systems such as fingerprints, voice and face since they are specific to an individual.

Advantages of biometrics: Cannot be forgotten or easily stolen.

Focus on human speech: Least obtrusive and easy to acquire.

2

Use of Smart Devices

- Smart devices technology: Significant increase.
- Smartphone users: 294 million in 2020.
- Smart home devices: 37% of households in the US.
- The smart devices with voice user interface needs a better user authentication and security since they interact with internet, controlling smart home devices and controlling another smart devices.

3

Speech Processing

- Applications: Speech recognition, language identification, and speaker recognition.
- Speaker recognition: Text-dependent or text-independent.
- Tasks:
  - Speaker Identification.
  - Speaker Detection.
  - Speaker Diarization.
  - **Speaker Verification.**

4

Automatic Speaker Verification

- Definition: Use of a machine to verify identity from voice.
- Interest in biometrics: Access-control applications.
- Difference from speaker identification and speaker verification.

5

Evaluation and Benchmarks

- Benchmarks: Ruled by the Linguistic Data Consortium (LDC) and the National Institute for Standards and Technology (NIST).
- NIST Speaker Recognition Evaluation (SRE): Focus on text independent speaker recognition tasks.
- Dataset: Multiple telephone calls, which were audited for language, speaker identity and other features.
- Other Datasets: Vox-Celeb, LDC, YOHO, TIMIT, AHUMADA.

6

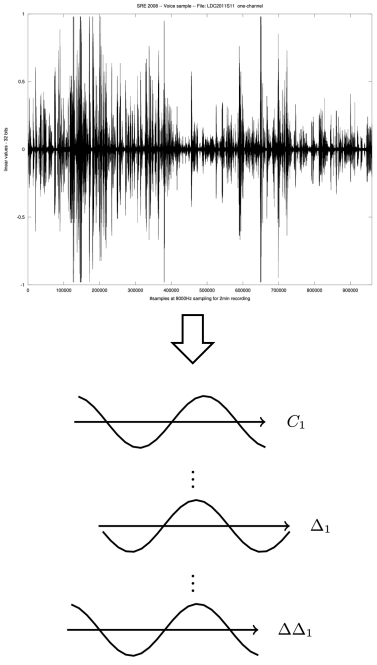
Voice Feature Extraction

- Definition: Transforming raw audio signals into measurable and meaningful characteristics.
- Goal: Capture essential information while reducing data size and eliminating noise.
- Mel-Frequency spaced Cepstral Coefficients (MFCCs):
  - Capture power spectrum.
  - Reflect human auditory perception.
- Other Common Feature Extractions: Perceptual Linear Prediction, Linear Predictive Coding, i-vectors.

7

Steps in Voice Feature Extraction

- Pre-processing:
  - Sampling: Convert audio to discrete-time signal.
  - Pre-emphasis: Emphasize higher frequencies.
- Framing:
  - Divide signal into overlapping frames.
- Windowing:
  - Apply window function to frames.
- Feature Calculation:
  - Time-Domain Features: Zero-crossing rate, energy, pitch.
  - Frequency-Domain Features: Fast Fourier Transform.
  - Cepstral Features: Mel-Frequency Cepstral Coefficients.
  - Delta and Delta-Delta Features: Capture dynamic changes.



8

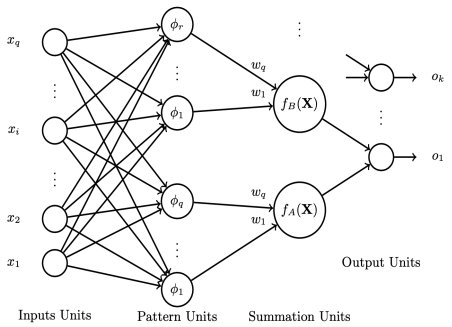
Models Used

- Probabilistic Neural Network (PNN): Inspired by Bayesian networks and uses kernel-based functions
- Key Characteristics:
  - Radial Basis Function (RBF): Uses Gaussian kernel functions to measure distances.
  - Spread Parameter (σ): Controls the width of the Gaussian functions.
  - Class Assignment: Based on normalized linear combinations of Gaussian outputs.

9

PNN Architecture

- Input Layer:
  - Accepts feature vectors (e.g., MFCCs from voice data).
- Pattern Layer:
  - Applies a Gaussian function to each input feature vector.
  - Basis functions (ϕi) measure similarity between input and stored patterns.
- Summation Layer:
  - Aggregates outputs from the pattern layer.
  - Computes weighted sums for each class.
- Output Layer:
  - Determines the class with the highest probability.
  - Final decision on classification.



$$\phi_i = exp\left(\frac{-\|\vec{X}-x_i\|^2}{2\sigma^2}\right)$$

10

ECAPA-TDNN

- Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Networks.
- Definition: A state-of-the-art neural network architecture designed for speaker verification.
- Key Characteristics:
  - Enhanced Performance: Better feature extraction and classification.
  - Contextual Awareness: Incorporates time-context to capture dynamic speaker features.
  - Advanced Aggregation: Uses sophisticated methods to combine feature representations.

11

Architecture

- Channel Attention Mechanism:
  - Emphasizes relevant acoustic features.
  - Enhances model's ability to focus on important speaker-specific cues.
- Information Propagation:
  - Efficiently transmits relevant information through the network.
  - Captures essential speaker characteristics.
- Aggregation Methods:
  - Combines information from different parts of the network.
  - Improves overall performance and robustness.

12

Data and Methodology

Dataset: NIST-SRE 2008.

Full dataset: 1,270 speakers and 98,777 test segments.

Custom dataset: 9 speakers and 111 test segments.

Speaker	Model Id	#test cases	#targets
1	10371	11	2
2	12459	10	1
3	12499	19	1
4	13076	14	4
5	13845	15	1
6	14692	10	4
7	14856	14	5
8	15604	9	1
9	16295	9	2
	Total	111	21

Model Id	Sex	Segment Id: Channel	Speaker Id	Speech type	Channel type	Lang.	Speaker native language
10371	m	tljiv:a	105088	phonecall	phn-main	ENG	USE
12459	m	ttqco:a	110372	phonecall	phn-main	ENG	USE
12499	m	trbfl:a	110667	phonecall	phn-main	ENG	USE
13076	m	tahak:b	103394	phonecall	phn-main	ENG	USE
13845	m	tbjem:a	111590	phonecall	phn-main	ENG	USE
14692	m	txkzs:b	103675	phonecall	phn-main	ENG	USE
14856	m	txqmx:b	102578	phonecall	phn-main	ENG	USE
15604	m	ttwkz:b	107851	phonecall	phn-main	ENG	USE
16295	m	tlskl:b	104179	phonecall	phn-main	ENG	USE

13

Equal Error Rate (EER)

- Definition: A performance metric used to evaluate biometric systems.
- False Acceptance Rate (FAR): The rate at which impostors are incorrectly accepted.
- False Rejection Rate (FRR): The rate at which legitimate users are incorrectly rejected.
- EER: The point at which FAR and FRR are equal. It provides a single value indicating the overall accuracy of the system.

14

Experimental Results PNN

- Spread parameter σ = 0.1.
- EER Results: Average 13.3%.

Speaker	Model Id	EER (%)	threshold	Accuracy
1	10371	0	0.78	1.00
2	12459	33.3	0.82	0.82
3	12499	22.2	0.78	0.78
4	13076	0	0.74	1.00
5	13845	35.7	0.84	0.66
6	14692	0	0.70	1.00
7	14856	0	0.78	1.00
8	15604	0	0.77	1.00
9	16295	28.5	0.82	0.77
	Average	13.3		0.88

15

Experimental Results ECAPA-TDNN

- Comparison with Speechbrain default configurations.
- EER Results Average: 41.39%

Speaker	Model Id	EER (%)	Threshold	Accuracy
1	10371	34.31	0.20	0.70
2	12459	41.24	0.06	0.57
3	12499	0.00	0.23	0.99
4	13076	40.59	0.19	0.55
5	13845	34.03	0.02	0.68
6	14692	58.13	0.17	0.48
7	14856	40.00	0.22	0.55
8	15604	51.67	0.16	0.53
9	16295	69.64	0.27	0.39
	Average	41.39		0.60

16

Conclusions

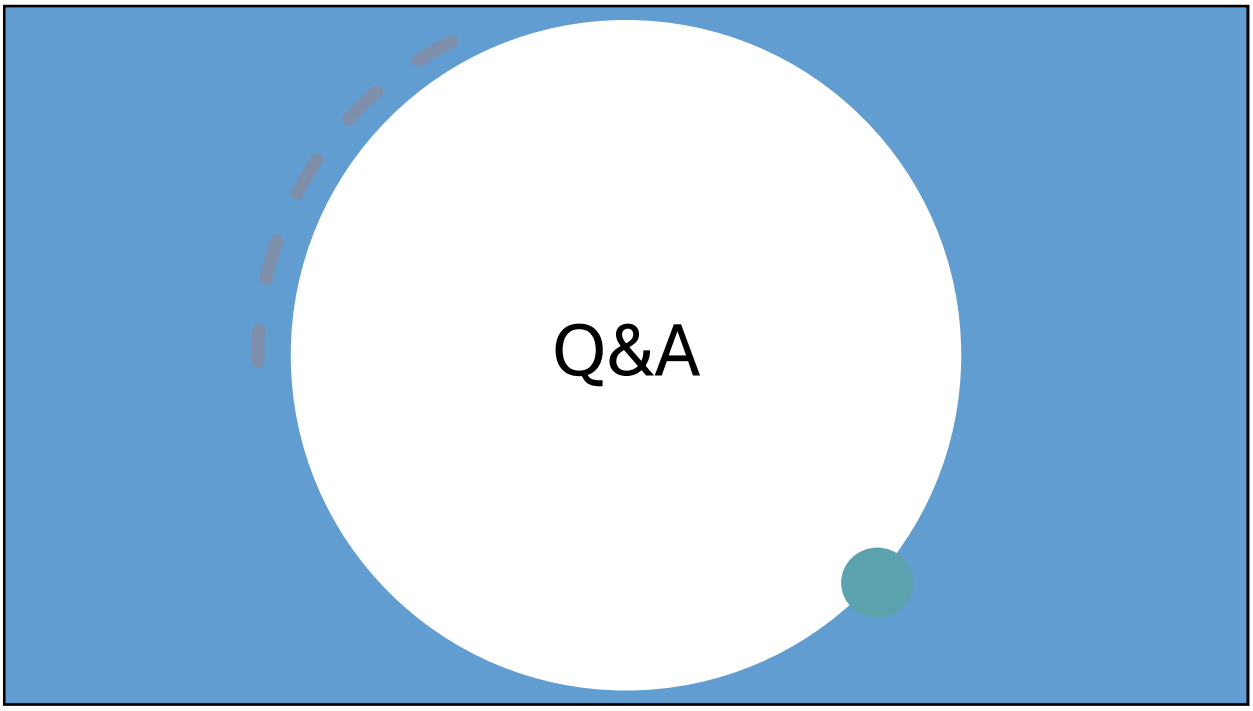
- PNN Performance: Best performance with EER of 0% for five out of nine speakers.
- ECAPA-TDNN: Expected improvements with larger datasets.
- Conclusion: PNN demonstrates superiority with small datasets.

17

Future Work

- Future Directions: Using the full SRE-08 dataset.
- Parallel Computing: Needed for PNN.
- Parameter Optimization: Adjusting spread parameter σ individually.
- Exploration of New Techniques: LPCs, PLP, i-vectors, mix PNN and ECAPA-TDNN.

18



19