УДК 530.1, 681.3.06

Janybekova S.T.[1], Tolganbayeva G.A.[2], Sarsembayev A.A.[3]

[1,2,3] International Information Technology University, Almaty, Kazakhstan

SPEAKER RECOGNITION USING DEEP LEARNING

Abstract. This paper discusses a transition from the traditional methods to novel deep learning architectures for speaker recognition. The article aims to compare the traditional statistical methods and new approaches using deep learning models. To articulate the difference in the discussed approaches it furthermore describes several recent methods of optimization and evaluation techniques. The review covers datasets used, results, contributions made toward speaker recognition, and limitations related to it.

Keywords: Speaker recognition, convolutional neural network, deep neural network, voice identification, recognition systems.

Introduction

Speaker recognition is the process of voice identification among a given set of speakers from a speech signal. Speech signal conveys information about the speaker's physiological properties [2], as well as behavioral aspects like accent and involuntary transforms of acoustic parameters [1] and several widely known application domains of speaker recognition. A user authentication in the bank sector is one of these applications. In February 2016 UK high-street bank HSBC [11] and its internet-based retail bank First Direct announced that it will offer its biometric banking software to access online and phone accounts using their fingerprint or voice [4] to their 15 million customers. Another application domain of growing popularity is home assistance.

The field of speaker recognition can be divided into speaker identification and speaker verification. It may be either open-set or closed-set [5]. The aim of speaker identification is to determine whether the voice of an unknown speaker matches one of the other speakers in a dataset, where the number of speakers could be very large. Speaker verification, on the other hand, is the process of determining who is speaking from known voices in the database. If the population of recorded voices is fixed, it is an open set. In contrast, closed-set verification is a case when new recordings of people can be added without having to redesign the system [6].

There are two types of systems commonly used in speaker recognition: text-dependent and text-independent. Text-dependent systems understand the content of the speech. Usually, the content is short utterances. In text-independent systems, there is no restriction on the spoken text. Forensic speaker ID is an example of text-independent applications [7].

Later in this article, we will discuss common traditional systems of speaker recognition such as GMM-UBM, GMM-SVM, and their limitations, as well as the use of deep learning technologies that have significantly advanced speaker recognition performance.

Feature extraction

The speech sound is a time-variant expressing various types of information, including text, speaker identities, acoustic features, emotions, etc. Speech may be observed in the time domain and frequency domain [4]. The most commonly used tool for visualizing speech is a spectrogram which describes the frequency spectra of consecutive short-term speech segments as an image. In the image, the horizontal and vertical axes represent time and frequency. The concentration of each point in the image is the magnitude of a distinct frequency and time. But for statistical modeling, spectrograms are not the best way [4]. One of the reasons is that the frequency dimension is too high. For the 1024-point fast Fourier transform, the frequency dimension is 512. Another reason is high correlation of frequency components with each other after FFT. A more compact illustration of

speech is obtained by using cepstral representation. Mel-frequency cepstral coefficients (MFCCs) are used to get features of speech. Figure 1.1 shows the technique of extracting MFCCs from a frame of the speech signal. In the figure, s(n) corresponds to a frame of speech, X(m) is the logarithm of the spectrum at frequencies determined by the *m*th filter in the filter bank, and MFCCs.

$$o_i = \sum_{m=1}^{M} \cos\left[i\left(m - \frac{1}{2}\frac{\pi}{M}\right)\right] X(m), i = 1\ldots, P \quad (1.1)$$

$$e = [e, o_1., o_p, \Delta e, \Delta o_1.., \Delta o_p, \Delta\Delta e, \Delta\Delta o_1.. \Delta\Delta o_p]^T \quad (1.2)$$

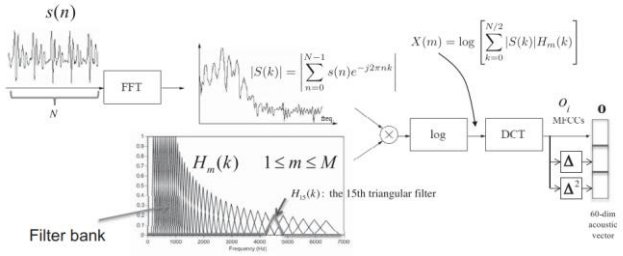

Figure 1.1 - Procedure of extracting MFCCs from a frame of speech. Refer to Eq. 1.1 and Eq. 1.2 for o_i and o respectively [4]

In Figure 1.2, the symbols Δ and $\Delta\Delta$ represent the velocity and acceleration of MFCCs, respectively. Denoted as e is the log-energy, an acoustic vector corresponding to $s(n)$.

One of the widespread approaches is the use of short-term spectral features: Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), linear predictive cepstral coefficients [1]. They are used due to their high performance and relatively low computational complexity [1].

Some recent DNN works suggest speaker recognition from raw waveforms using Convolutional Neural Networks (CNNs). Instead of using standard hand-crafted features, networks learn low-level speech representations from raw waveforms. That allows to better get significant narrow-band characteristics such as pitch and formats. ResNet-base, VGG-M based trunk CNN architectures are used for spectrogram inputs [5].

Probabilistic Models

One category of statistical learning methods is distinguished as latent variable models that intend to relate a set of observable variables X to a set of latent variables Z based on the number of probability distributions. Λ parameters of the probability functions are evaluated by maximizing the likelihood function with respect to model parameters [5].

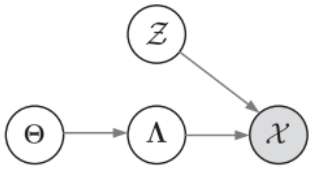

Figure 2.1 - A hierarchical Bayesian representation for a latent variable model with parameters, latent variables Z, observed variables χ and hyperparameters. The shaded node denotes observations. The unshaded nodes mean latent variables [5]

Based upon the hyperparameters , model parameters are represented by a prior density $p(\Lambda|\Theta)$. Having the model parameters, the speech training samples χ are generated by a likelihood

function $p(\chi|\ \Lambda)$, which is marginalized over discrete latent variables by

$$p(\chi\ |\Lambda)\ =\ \Sigma_z\quad\ p\ (\chi|\Lambda) \tag{1.3}$$

through a continuous latent variable by

$$p(\chi\ |\Lambda)\ =\ \int\quad\ p\,(\chi, Z\ |\ \Lambda)\ dZ \tag{1.4}$$

The conditional likelihood with discrete latent variables is expressed by

$$p\ (Y|\ \chi,\Lambda)\ =\ \Sigma_z\quad p(\ Y,\Lambda)\ =\ \Sigma_z\quad\ p(Y|\chi,\Lambda).p(Z|\Lambda) \tag{1.5}$$

Unsupervised, supervised, and semi-supervised models can be adaptably performed and constructed by optimizing the corresponding likelihood functions in an individual or hybrid style [5].

The probabilistic methods are complex for real-world applications. There is a variety of approximate inference algorithms that can solve the optimization problem, but they are generally hard to derive. Indirect optimization over the evidence lower bound (ELBO) is implemented as an analytical solution [5]. There are different latent variable models such as Gaussian Mixture models (GMM), joint factor analysis (JFA), probabilistic linear discriminant analysis (PLDA), factor analysis (FA), a mixture of PLDA. The GMM-UBM system is a direct generative approach for speaker verification tasks. In this method, the training phase is preceded by the estimation of s speaker-independent universal background model (UBM), using a large voice data of several hours. The UBM is where

$$\lambda\ _{UBM}\ =\ \{w_i,\mu\ _i,\Sigma\ _i\}\ _{i=1}^{C} \tag{1.6}$$

C is the quantity of Gaussian components, $\omega\ _i$ is the prior of i-th Gaussian component, $\mu\ _i$ is mean and $\Sigma\ _i$ - covariance matrix. Each speaker is an adaptation from UBM.

i-vector system is the high dimensional GMM supervector in a total variability (TV) space. It reduces the supervector into low dimensional factors [1]. Concatenated means of GMM are presented as

$$M\ =\ m\ +\ \Phi y \tag{1.7}$$

where Φ is a low-rank factor loading matrix and m - channel. $M$is a supervector of speech utterance with feature vectors.

Deep Neural Networks

Deep Neural Networks are trained to discriminate between speakers, map variable-length utterances to fixed-dimensional embeddings that are called x-vectors. Although most speaker recognition systems are based on i-vectors, x-vectors used like i-vectors bit built on DNN embedding architectures [6] are used in novel publications. For visual comparison there are i-vector and x-vector pipelines shown in Figure 3.1

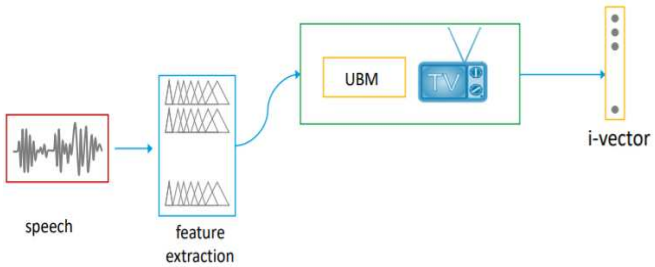

Figure 3.1 - The i-vector pipeline [7]

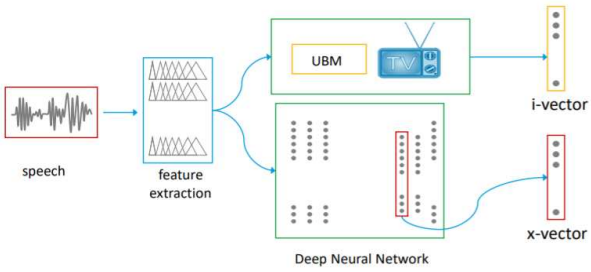

Figure 3.2 - The i-vector and x-vector pipelines [7]

The x-vector system is based on a framework for speaker recognition. The system is composed of a feed-forward deep neural network that maps variable-length speech segments to embeddings that are called x-vectors [8]. Those vectors are classified by trained Gaussian classifiers. The network is implemented using the nnet3 neural network library in the Kaldi Speech Recognition Toolkit [3]. The recipe is based on the SRE16 v2 recipe available in the main branch of Kaldi [8].

The *d*-vector was developed using multiple fully-connected neural network layers and *X*-vectors which are based on the Time-delayed neural networks (TDNN) that are popular in recent years [7].

Evaluation metrics

For closed-set speaker recognition, accuracy (recognition rate) is the usual performance measure [4].

$$Recognition\ rate = \frac{\#\ of\ correct\ recognition}{Total\ \#\ of\ trials} \tag{1.8}$$

Other popular measures include the false rejection rate (FRR), the actual decision cost function (DCF), the minimum decision cost function (min DCF), and the equal error rate (EER). Their principles are very similar.

The definition of FAR and FRR are the following:

$$False\ reject\ rate\ (FRR) = Miss\ probability = \frac{\#\ of\ true-speakers\ rejected}{Total\ \#.of\ true-speaker\ trials} \tag{1.9}$$

$$False\ acceptance\ rate\ (FAR) = \frac{\#\ of\ impostors\ accepted}{Total\ \#\ of\ impostors\ attempts} \tag{2.1}$$

Concepts of FAR, FRR, ERR are explained in Figure 4.1. They use the distributions of speaker scores and impostor scores of two speaker verification systems which are System A and System B.

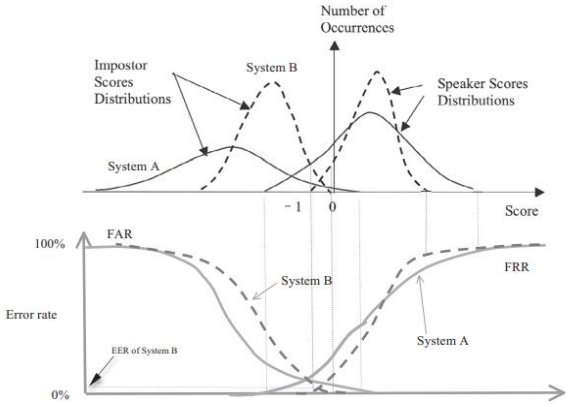

Figure 4.1 - Distributions of true speaker scores and impostor scores of two speaker verification systems. EER, FAR, FRR of two systems [4]

Some results of the comparison and advancements on i-vector based and x-vectors speaker recognition techniques are presented in Table 1.4 for better understanding.

The primary performance measure for the Conversational Telephone Speech (CTS) Speaker Recognition Challenge (CTS SRE) in 2019 was a detection cost described as a weighted sum of false-reject (miss) and false-accept (false alarm) error probabilities. The CTS Challenge primarily normalized the cost function for a decision threshold θ [3]

$$C_{norm}(\Theta) = P_{miss}(\Theta) + \beta \times P_{fa}(\Theta), \quad (2.2)$$

where $\beta$is

$$\beta = \frac{C_{fa}}{C_{miss}} \times \frac{1 - P_{target}}{P_{target}} \quad (2.3)$$

where C_{min} - is the cost of a missed detection and C_{fa} - the cost of a false alarm. $P_{target}$is the a priori probability that the test segment speaker is the specified target speaker. The primary cost metric, $C_{primary}$ for the CTS Challenge was the average of normalized costs calculated at two points along the detection error trade-off (DET) curve [3], with C_{miss}=C_{fa}= 1, P_{target} = 0.01, and P_{target} = 0.005.

Table 1.1 Recent results of EER i-vectors and DNN methods

Year	Datasets	Evaluation Results	
2018 Results for VoxCeleb 1 verification [9]	VoxCeleb 1 consists of over 100,000 utterances for 1,251 celebrities, extracted from videos uploaded to YouTube [9].	GMM-UBM	15.0
		I-vectors + PLDA	8.8
		CNN - 1024	10.2
		CNN + Embedding	7.8
2018 Results for verification on the original VoxCeleb2 test set [10]	Test VoxCeleb2 VoxCeleb2 consists of over 1 million utterances for over 6,000 celebrities. The dataset is reasonably gender-balanced, where 61% of the speakers are males	VGG-M	5.94
		ResNet-34	4.83
		ResNet-50	3.95
2020 Verification performance on full utterance (NS: Normalized softmax; TAP: Temporal Average Pooling) [12]	VoxCeleb 1	ResNet34 (TAP+NS)	3.81
	VoxCeleb 2	ResNet34 (TAP+NS)	2.08
2020 E-TDNN (Densely Connected Time Delay Neural Network for	VoxCeleb1 test set	E-TDNN	4.65

Speaker Verification) [13]			
2020 Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs [13]	VoxCeleb1 test set	D-TDNN-SS	1.22

Conclusion

This paper has provided a brief review of Probabilistic Models and deep learning techniques for speaker recognition. Deep learning techniques such as CNN have been the subject of much research in recent years. This research considers limitations of traditional techniques and forms a base to evaluate the performance and limitations of the current deep learning techniques. Further, it highlights some promising directions for better speaker recognition systems. It is still a non-trivial task when recognizing or verifying speakers in poor acoustic conditions. The paper also compares several model deep learning approaches for speaker recognition tasks, their evaluation metrics and datasets.

REFERENCES

1. Poddar, A., Sahidullah, M. and Saha, G., 2017. Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics, 7(2), pp.91-101.
2. Kinnunen, T. and Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), pp.12-40.
3. Sadjadi, S.O., Greenberg, C., Singer, E., Reynolds, D., Mason, L. and Hernandez-Cordero, J., 2020, May. The 2019 NIST speaker recognition evaluation CTS challenge. In Speaker Odyssey (Vol. 2020, pp. 266-272).
4. Mak, M.W. and Chien, J.T., 2020. Machine learning for speaker recognition. Cambridge University Press pp. 3-72.
5. Ravanelli, M. and Bengio, Y., 2018, December. Speaker recognition from raw waveform with sincnet. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 1021-1028). IEEE.
6. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S., 2018, April. X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329-5333). IEEE.
7. Kelly, F., Alexander, A., Forth, O. and van der Vloed, D., From i-vectors to x-vectors–a generational change in speaker recognition illustrated on the NFI-FRIDA database, pp. 1-5.
8. Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D. and Khudanpur, S., 2018, June. Spoken language recognition using x-vectors. In Odyssey (pp. 105-111).
9. Nagrani, A., Chung, J.S. and Zisserman, A., 2017. Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612. (pp. 1-3).
10. Chung, J.S., Nagrani, A. and Zisserman, A., 2018. Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (pp. 1-5).
11. HSBC Holdings plc is a British multinational investment bank and financial services holding company, https://en.wikipedia.org/wiki/HSBC.
12. Kye, S.M., Jung, Y., Lee, H.B., Hwang, S.J. and Kim, H., 2020. Meta-learning for short utterance speaker recognition with imbalance length pairs. arXiv preprint arXiv:2004.02863 (pp. 1-5).
13. Yu, Y.Q. and Li, W.J., 2020. Densely Connected Time Delay Neural Network for Speaker Verification. Proc. Interspeech 2020, (pp.921-925)

Джаныбекова С.Т., Толғанбаева Г.А., Сарсембаев А.А.
Распознавание говорящего с помощью глубокого обучения

Аннотация: В этой статье обсуждается переход от традиционных методов к новым архитектурам глубокого обучения для распознавания говорящего. Он направлен на сравнение традиционных статистических методов и новых подходов с использованием моделей глубокого обучения. Также описаны новейшие методы оптимизации. Из-за разных подходов существует несколько методик оценки. В этой статье представлен обзор методов глубокого обучения и обсуждается недавняя литература, в которой эти методы используются для распознавания речи. Обзор охватывает используемые базы данных, результаты, вклад в распознавание речи и связанные с этим ограничения.

Ключевые слова: распознавание говорящего, сверточная нейронная сеть, глубокая нейронная сеть, идентификация по голосу, системы распознавания.

Джаныбекова С.Т., Толғанбаева Г.А., Сарсембаев А.А.
Терең оқыту арқылы сөйлеушіні тану

Аңдатпа: Бұл мақалада сөйлеушілерді тану үшін дәстүрлі әдістерден жаңа терең оқыту архитектурасына көшу туралы айтылады. Ол тереңдетілген оқыту модельдерін қолдана отырып дәстүрлі статистикалық әдістер мен жаңа тәсілдерді салыстыруға бағытталған. Сонымен қатар оңтайландырудың соңғы әдістері сипатталған. Сондай-ақ әртүрлі тәсілдерге байланысты бағалаудың бірнеше әдістемесі бар. Бұл мақалада терең оқыту әдістеріне шолу жасалады және сөйлеуді тану үшін осы тәсілдерді қолданатын соңғы әдебиеттер талқыланады. Шолу пайдаланылған мәлімет базасын, нәтижелерді, сөйлеуді тануға қосқан үлестерін және осыған байланысты шектеулерді қамтиды.

Түйінді сөздер: сөйлеушіні тану, конволюциялық жүйке жүйесі, терең жүйке жүйесі, дауысты сәйкестендіру, тану жүйелері.

Авторлар туралы мәлімет:

Сарсембаев Айдос Айдарович, PhD «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының ассистенті, Халықаралық ақпараттық технологиялар университеті.

Джаныбекова Салтанат Талгатбековна, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының докторанты, Халықаралық ақпараттық технологиялар университеті.

Толғанбаева Гауһартас Алғабасқызы «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының докторанты**,** Халықаралық ақпараттық технологиялар университеті.

Сведения об авторах:

Сарсембаев Айдос Айдарович, PhD, ассистент-профессор кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

Джаныбекова Салтанат Талгатбековна, докторант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

Толғанбаева Гауһартас Алғабасқызы, докторант кафедры «Компьютерная инженерия и информационная безопасность», Международный университет информационных технологий.

About the authors:

Aidos A. Sarsembayev, Ph.D., Assistant-Professor, Department of Computer Engineering and Information Security, International Information Technology University.

Saltanat T. Janybekova doctoral student, Department of Computer Engineering and Information Security, International Information Technology University.

Gaukhartas A. Tolganbayeva, doctoral student, Department of Computer Engineering and Information Security, International Information Technology University.