# Voice Authentication Model for One-time Password Using Deep Learning Models

Bella
Universitas Tarumanagara
Letjen S. Parman St. No.1
DKI Jakarta, Indonesia
mz.bellalie@gmail.com

Janson Hendryli
Universitas Tarumanagara
Letjen S. Parman St. No.1
DKI Jakarta, Indonesia
jansonh@fti.untar.ac.id

Dyah Erny Herwindiati
Universitas Tarumanagara
Letjen S. Parman St. No.1
DKI Jakarta, Indonesia
dyahh@fti.untar.ac.id

## ABSTRACT
This paper explores the possibility of implementing a voice authentication system consisting of speech recognition and speaker verification model for the one-time password (OTP) system. The speech recognition model is responsible for classifying user utterances of random OTP digits in Bahasa Indonesia and the speaker verification model is used to verify the identity of the speaker. The long short-term memory network and siamese network with convolutional neural networks are employed as the model, where they aim to recognize and verify human voices represented by MFCC feature vectors. From the experiments, it is found that the validation accuracy of the speech recognition model is reliable, yet the speaker verification model cannot achieve satisfactory result.

## CCS Concepts
**• Security and privacy→Authentication • Computing methodologies→Artificial intelligence • Supervised learning • Neural networks • Speech recognition**

## Keywords
Speech recognition; speaker verification; deep learning; one-time password

## 1. INTRODUCTION
All applications, either web-based or mobile-based, nowadays must have a secure authentication system. Authentication itself is the process of verifying the identity of the user by obtaining some sort of credentials and, using those credentials, verify the user's identity. With the increasing case of fraud cases, there may be an alternative of credential information to verify the identity of a user. Hence, the two factor authentication system is introduced.

Many applications have implement two factor authentication system, e.g. using additional password and one-time password. One-time password, or frequently abbreviated as OTP, is a random generated single-use password in the form of numbers to ensure that the rightful user is the person who attempts the log in or transaction process. The OTP's advantage over traditional password or PIN numbers is that the numbers are always randomly generated and can only be used once.


*BDET2020*, January 3–5, 2020, Singapore.

ACM ISBN 978-1-4503-7683-9/20/01…$15.00
DOI: https://doi.org/10.1145/3378904.3378908

This paper explores the models for building an authentication system using human voice. Human voice has unique characteristic, which is different from one person to another, such that the user can use their own voice to authenticate their identity. In addition to that, to prevent voice recording, the system generates random numbers as the OTP and the authenticated user has to say the exact same random number. The proposed authentication system uses human voices as the authentication method where the user records their voice saying five digits of random number in Bahasa Indonesia. The system will then authenticate the transactions by recognizing the digits and verifying the speaker.

To have the ability of verifying and recognizing the utterances, the voice authentication system consists of two parts, which are the speaker verification model and speech recognition model. The speaker verification model is used in the process of verifying the user identity based on the speaker's registered voices. Meanwhile, the speech recognition system classifies the user utterances into nine digits of numbers (one to nine) in Bahasa Indonesia. The uttered digits have to be identical to the random generated digits. It should be noted that the zero digit in Indonesian is excluded from the system, because there are many variation in pronunciation of zero in Bahasa Indonesia, e.g."*nol*" and"*kosong*".

Several past research have introduced the text-dependent speaker verification model using end-to-end method with deep neural networks and long short-term memory networks [7]. In [17], the author also explores a text-dependent speaker verification using siamese network and sequence-to-sequence model with attention. For the speech recognition task, some noteworthy research are [13] where they explore speech recog-nition model for noisy data using very deep convolutional neural networks, [6] who are studying speech recognition using deep bidirectional LSTM and connectionist temporal classification, and in [9] where they explore frequency-time LSTM for speech recognition.

For building the voice verification model in this research, we propose the siamese network consisting of two identical convolutional neural networks. Moreover, the long short-term memory networks are utilized for the speech recognition model. As for the feature representation of human voices, we use the Mel frequency cepstral coefficients.

## 2. METHODS
### 2.1 Mel Frequency Cepstral Coefficients
Mel frequency cepstral coefficients, known as MFCC, is one of the most popular features representation in auto-matic speech recognition. MFCC is also used as features in several research, such as for gender recognition [12], speech recognition [1], and

speaker verification [5]. The input for the MFCC extraction process is raw audio files in WAV format, and the output is the sample of frames and cep-stral coefficients. The steps in the calculation of MFCC are pre-emphasis, framing, windowing, calculating fast Fourier transform, applying mel filter bank, and finally, taking the discrete cosine transform of the mel log powers returns the amplitudes of the resulting spectrum as the MFCC features. The detail of each processes in MFCC is explained in [12].

## 2.2 Deep Learning Models

**Convolutional neural network**, or more well known as CNN or convnet, is one of deep learning models which commonly used for image classification. Several main operations in CNN are the convolution process, pooling, and fully connected layer. Batch normalization layer is also frequently used in many research. While CNN is originally used for image classification task and uses 2-dimensional convolutional operation, it can also be used for sequential data by using the 1-dimensional convolutions. For example, 1-dimensional CNN which is used for speaker identification is studied in [5].

**Siamese network** is a network consisting of two identical networks sharing weights with the objective is to learn a similarity function [4]. Siamese network is used for one-shot learning task for face verification in [16]. Typically, siamese network uses two identical CNNs for learning the feature representation and measures how similar the two inputs are. This problem can be addressed as binary classification, because the main objective is to identify whether two input voices belong to the same person or not.

**Long short-term memory network**, commonly known as LSTM network, is one of popular methods used for learning from sequential data, such as for speech recognition [15]. LSTM is originally introduced by [8] to overcome the long-term dependencies problem in the original recurrent neural network. The LSTM consists of four gates to control the information to be learned, which are the input gate, forget gate, candidate gate, and output gate [14]. Also, there are two states, cell state and hidden state, which hold information to be passed from each LSTM cells in one time step to another [14]. LSTM has also been explored in [10], where the authors focus on solving large vocabulary speech recognition using LSTM, and in [3] for speech enhancement and automatic speech recognition.

## 2.3 Youden's Index

Youden's index is used in the speaker verification system to compute the optimal threshold for accepting the user's voices only if the speaker is the same as the user who he/she claim to be. The optimal threshold is obtained by getting the maximum Youden's index value for all possible threshold. Youden's index (J) is describe as in Eq. 1 [2]:

$$J = \text{sensitivity} + \text{specificity} - 1 \quad (1)$$

where sensitivity and specificity can be computed as in Eq. 2 and Eq. 3, respectively.

$$\text{sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (2)$$

$$\text{specificity} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}} \quad (3)$$

# 3. EXPERIMENTS

## 3.1 Data

The voice data for the speaker verification and speech recognition are collected from the respondents in both sexes who record their voices saying digit number from one to nine in Bahasa Indonesia. All of the respondents lives in Jakarta, Indonesia. The recordings are all from 8 female and 12 male respondents. The recordings for each digits are taken in WAV format within duration of one to two seconds. For every respondents, each digit numbers is recorded repeatedly for 5 times. In addition to that, although the recordings are taken in a quiet place, there are still some background noises, such as the sound of the surrounding environment and air conditioning system.

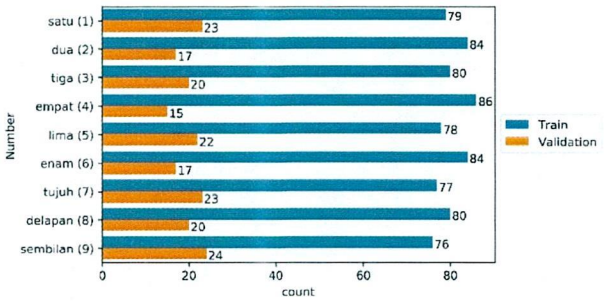


**Figure 1. Data for the training and validation of speech recognition model**

For the speech recognition task, the data set consists of 905 recordings from 20 respondents. We split the data proportionally for the training and validation phase as in Fig. 1. Meanwhile, for training the speaker verification model, we need pairwise data which consist of positive and negative labels, where positive label denotes recordings of the same person and likewise, recordings that belong to two different people are labelled negative. Furthermore, each pairs of the recording data for training the speaker verification model is of the same digit number. To illustrate, a positive labelled pair or matching pair is of person A saying *satu* and another recording of him also saying *satu*. Accordingly, the pair of person A saying *dua* and person B saying *dua* is labelled as negative or non-matching pair.

From 899 recordings of 23 respondents, we split the data into 539 recordings for training the speaker verification model and 360 recordings for the validation stage. Creating positive and negative pairs results in more negative labels than the positive ones. So in order to balance the pairings, we also under sample the negative pairs.

## 3.2 Implementation

The sampling rate of the voice recordings is 22050 Hz. In the preprocessing step, we perform the silence removal process at the beginning and end of every recordings. Since each recordings has different duration and signal length, before the feature extraction process begins, recordings with shorter length are zero-padded. To extract the MFCC features, we use the LibROSA [11] library in Python. The frame blocking step produces 125 frames, and with 40 filter banks, 13 cepstral coefficients are obtained.

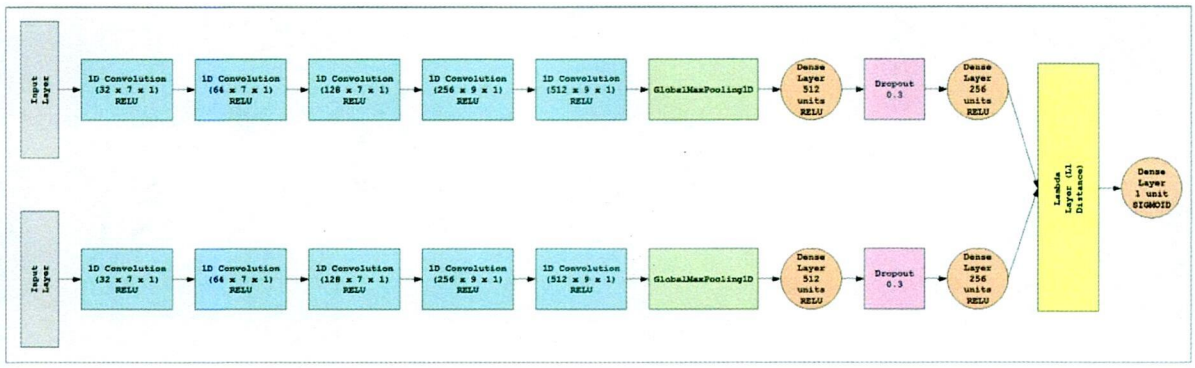


**Figure 2. Siamese Network with 2 CNN for Model SV1**

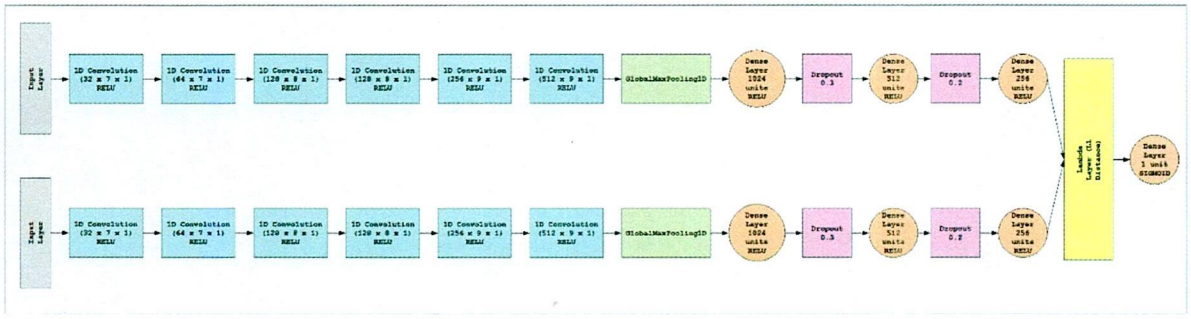


**Figure 3. Siamese Network with 2 CNN for Model SV2**

For the LSTM, the input of every time steps is one frame consisting of 13 cepstral coefficients. At the end of the time step, the output of the last hidden state is fed into a dense layer with 9 units and softmax activation function. The speech recognition model uses LSTM network with 256 units, followed by a fully connected layer with softmax activation function. The model was trained in 100 epochs, batch size of 8, and learning rate 0.0003051074046. To monitor the training process, early stopping with patience of 10 and model checkpoint to save the best weights are also employed. Furthermore, the models are trained using Adam optimizer and cross entropy as the loss function.

Fig. 2 and 3 show the siamese networks for the speaker verification model, which we subsequently call SV1 and SV2. The models are trained with Adam optimizers in 100 epochs with binary cross entropy loss function, batch size of 16, and also learning rate of 0.000027 for SV1 and 0.000017 for SV2.

Moreover, we compute the L1-distance as a measure of how similar the two inputs are. Similar to the training of speech recognition model, we also employ early stopping with patience of 10 and model checkpoint.

# 4. RESULTS

## 4.1 Speech Recognition

The speech recognition models are being evaluated for 9 digit classes. Fig. 4 shows the accuracy for the speech recognition models. The best training accuracy is achieved by the model where it stops early at 52$^{nd}$ epochs with 100% training accuracy and 96% validation accuracy. Moreover, the precision, recall, and F1-score of the speech recognition model on the validation data can be seen in Fig. 5.

## 4.2 Speaker Verification

During the training process, model SV1 and SV2 stop early at the 38$^{th}$ and 29$^{th}$ epoch, respectively. The SV1 model achieves 69% training accuracy and SV2 achieves 70% training accuracy. For the validation, both models achieve 61% in accuracy. It should be noted that from the validation data, we find the optimal threshold for categorizing positive or negative pairs using Youden's index. The threshold is presented in Table 1.

**Table 1. Optimal threshold for the speaker verification models**

| Models | J | TPR | FPR | Threshold |
|---|---|---|---|---|
| SV1 | 0.314136 | 0.607330 | 0.293194 | **0.510011** |
| SV2 | 0.287958 | 0.659686 | 0.371728 | **0.455436** |

To evaluate the speaker verification model better, Fig. 6 shows the ROC curve and the AUC score of both models. From the AUC score, we look further to the best performing model, which is the SV1 model, and plot the training accuracy as in Fig. 7. Furthermore, the precision, recall, and F1-score of SV1 model on validation data are shown as in Fig. 8 where 0 is the negative label and 1 is the positive label. It can be seen that the average F1-score of SV1 model is 65.5%.

## 4.3 Discussions

In the speech recognition task, the F1-score are all above 90%, especially for the digit *satu* (1), *dua* (2), and *tujuh* (7) where the model can perfectly recognize it. For the speaker verification task, we experiment on two models, SV1 and SV2, which have not achieved satisfying performance with less than 70% in accuracy. The AUC scores of the models are lower than 0.7, and also, the performance between SV1 and SV2 are not significantly different. The ROC curves of both models also indicate that both models

have low true positive rate and high false positive rate. It is important to be noted that speech recognition models and speaker verification models which has been trained has not been tested with the unseen data (excluded from the training and validation data) because the models have not achieve good result.

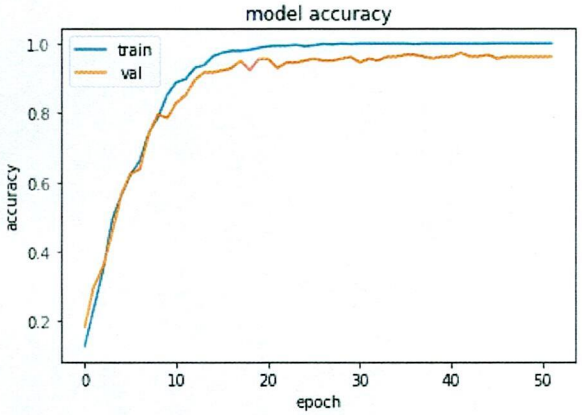


**Figure 4. The accuracy of the speech recognition model on training and validation data**

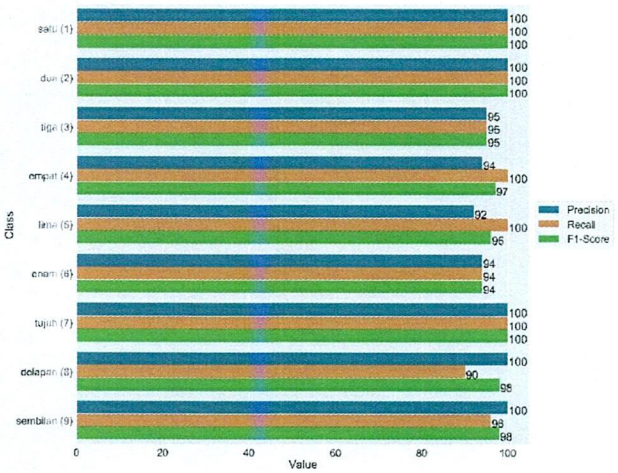


**Figure 5. The precision, recall, and F1-Score for each nine digits from testing SR3 model using validation data**

## 5. CONCLUSIONS

This paper proposes a voice authentication models which can be used in an OTP-like security system consisting of a speech recognition and speaker verification model. The speech recognition model is implemented using long short-term memory networks and the siamese network with convolutional neural networks is employed for the speaker verification model. From collected voice recording of people saying nine digits of numbers in Bahasa Indonesia, we find that the speech recognition model yields satisfactory performance. Meanwhile, for the speaker verification model, the accuracy is not good enough to build a reliable security sys-tem and needs to be explored. Improving the performance of the speaker verification model, such as adding more data and further tuning the hyperparameters and architecture of the convolutional neural networks, become our main concern for the future works.

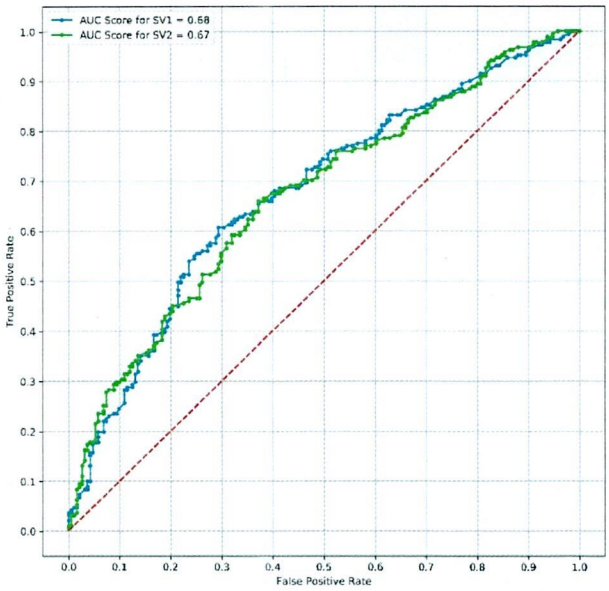


**Figure 6. The ROC Curve and AUC Score of speaker verification models**

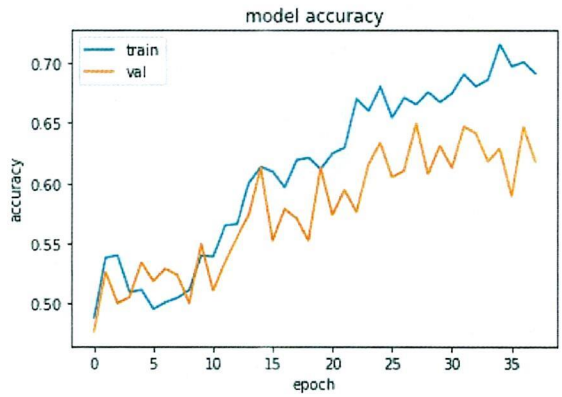


**Figure 7. The training and validation accuracy of SV1 model**

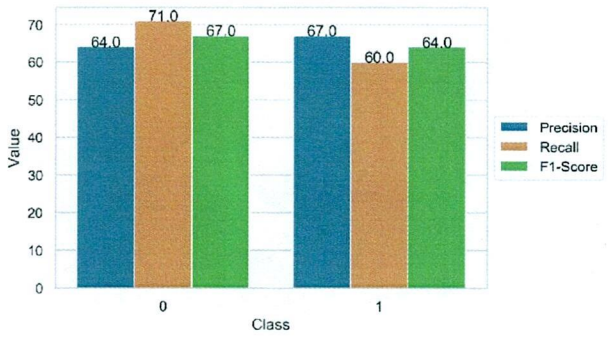


**Figure 8. The precision, recall, and F1-Score of SV1 model**

## 6. REFERENCES


[1] Aggarwal, N. 2015. Analysis of various features using different temporal derivatives from speech signals. *Int. J. Comput. Appl.* 118, 8 (May. 2015), 1-9.

[2] Carter, J. V., Pan, J., Rai, S. N., and Galandiuk, S. 2016. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. *Surgery* 159, 6 (Jun. 2016), 1638-1645. DOI=https://doi.org/10.1016/j.surg.2015.12.029.

[3] Chen, Z., Watanabe, S., Erdogan, H., and Hershey, J. R. 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (Dresden, Germany, September 6-10, 2015). INTERSPEECH '15. 3274-3278.

[4] Chopra, S., Hadsell, R., LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Diego, CA, United States, June 20-25, 2015). CVPR '05. IEEE, New York, NY, 539-546. DOI=https://doi.org/10.1109/CVPR.2005.202.

[5] Chowdhury, A. and Ross, A. 2017. Extracting sub-glottal and supra-glottal features from MFCC using convolutional neural networks for speaker identification in degraded audio signals. In *Proceedings of the 2017 IEEE International Joint Conference on Biometrics* (Denver, CO, United States, October 1-4, 2017). IJCB '17. IEEE, New York, NY, 608-617. DOI=https://doi.org/10.1109/BTAS.2017.8272748.

[6] Graves, A. and Jaitly, N. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning* (Beijing, China, June 21-26, 2014). ICML '14. JMLR, 1764-1772.

[7] Heigold, G., Moreno, I., Bengio, S. and Shazeer, N. 2016. End-to-end text-dependent speaker verification. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (Shanghai, China, March 20-25, 2016). ICASSP '16. IEEE, New York, NY, 5115-5119. DOI=https://doi.org/10.1109/ICASSP.2016.7472652.

[8] Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9, 8 (Nov. 1997), 1735-1780. DOI=https://doi.org/10.1162/neco.1997.9.8.1735.

[9] Li, J., Mohamed, A., Zweig, G., and Gong, Y. 2015. LSTM time and frequency recurrence for automatic speech recognition. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding* (Scottsdale, AZ, United States, December 13-17, 2015). ASRU '15. IEEE, New York, NY, 187-191. DOI=https://doi.org/10.1109/ASRU.2015.7404793.

[10] Li, X. and Wu, X. 2015. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (Brisbane, Australia, April 19-24, 2015). ICASSP '15. IEEE, New York, NY, 4520-4524. DOI=https://doi.org/10.1109/ICASSP.2015.7178826.

[11] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. 2015. Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (Austin, Texas, July 6-12, 2015). SciPy '15. 18-24.

[12] Pahwa, A. and Aggarwal, G. 2016. Speech feature extraction for gender recognition. *Int. J. Image, Graph. and Signal Processing* 8, 9 (Sept. 2016), 17-25. DOI=https://doi.org/10.5815/ijigsp.2016.09.03.

[13] Qian, Y., Bi, M., Tan, T., and Yu, K. 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Trans. Audio, Speech, and Lang. Process.* 24, 12 (Dec. 2016), 2263-2276. DOI=https://doi.org/10.1109/TASLP.2016.2602884.

[14] Ravuri, S. and Stolcke, A. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (Dresden, Germany, September 6-10, 2015). INTERSPEECH '15. 135-139.

[15] Sainath, T. N., Vinyals, O., Senior, A., and Sak, H. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (Brisbane, Australia, April 19-24, 2015). ICASSP '15. IEEE, New York, NY, 4580-4584. DOI=https://doi.org/10.1109/ICASSP.2015.7178838.

[16] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH, United States, June 23-28, 2014). CVPR '14. IEEE, New York, NY, 1701-1708. DOI= http://dx.doi.org/10.1109/FCVPR.2014.220.

[17] Zhang, Y., Yu, M., Li, N., Yu, C., Cui, J., and Yu, D. 2019. Seq2Seq attentional siamese neural networks for text-dependent speaker verification. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (Brighton, United Kingdom, May 12-17, 2019). ICASSP '19. IEEE, New York, NY, 6131–6135. DOI=https://doi.org/10.1109/ICASSP.2019.8682676.