Robust Anchorperson Detection Based on Audio Streams using a Hybrid I-vector and DNN System

Yun-Fan Chang[*], Payton Lin[*], Shao-Hua Cheng[†], Kai-Hsuan Chan[†], Yi-Chong Zeng[†], Chia-Wei Liao[†], Wen-Tsung Chang[†], Yu-Chiang Wang[*], Yu Tsao[*]
[*]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[†] Advanced Research Institute, Institute for Information Industry, Taipei, Taiwan

***Abstract*— Anchorperson segment detection enables efficient video content indexing for information retrieval. Anchorperson detection based on audio analysis has gained popularity due to lower computational complexity and satisfactory performance. This paper presents a robust framework using a hybrid I-vector and deep neural network (DNN) system to perform anchorperson detection based on audio streams of video content. The proposed system first applies I-vector to extract speaker identity features from the audio data. With the extracted speaker identity features, a DNN classifier is then used to verify the claimed anchorperson identity. In addition, subspace feature normalization (SFN) is incorporated into the hybrid system for robust feature extraction to compensate the audio mismatch issues caused by recording devices. An anchorperson verification experiment was conducted to evaluate the equal error rate (EER) of the proposed hybrid system. Experimental results demonstrate that the proposed system outperforms the state-of-the-art hybrid I-vector and support vector machine (SVM) system. Moreover, the proposed system was further enhanced by integrating SFN to effectively compensate the audio mismatch issues in anchorperson detection tasks.**

I. INTRODUCTION

Efficient browsing and retrieval processes in digitally recorded video has become a critical task due to the rapid expansion of large video databases. Anchorperson detection has served as an effective method to segment video for facilitating subsequent information retrieval processes for particular programs such as broadcast news, sports, and talk shows. Recently, anchorperson detection using audio parts of video data has gained interest due to relatively cheaper computation and reliable performance [1-3].

Anchorperson detection based on audio data can be accomplished by a single–target speaker and text-independent speaker verification system. Previous studies have proposed using Gaussian mixture model (GMM) to characterize speakers' voices [4-6]. These approaches first prepare a universal background model (UBM) GMM, and utilize speaker-specific spoken materials to adapt UBM-GMM to obtain speaker-specific models. In the testing set, the likelihood ratio of the UBM-GMM and speaker-specific GMM is computed to verify the speaker's claimed identity. To enhance the characterization capability of temporal information of speech signals, acoustic segment model (ASM) [7] and hidden Markov model (HMM) [8-10] have been developed for speaker models. These models can also be used as the first stage, followed by a discriminative classifier such as support vector machine (SVM) [11] and artificial neural network (ANN) [12] to perform classification. More recently, the I-vector approach has been developed and has become the state-of-the-art technique in speaker recognition tasks [13, 14]. In the I-vector approach, each utterance is represented by a single vector whose elements are essentially the latent variables of a factor analyzer. With the I-vector as the feature extraction stage, a follow-up SVM is applied to perform classification [13, 14].

The present study proposes a hybrid I-vector and deep neural network (DNN) speaker verification scenario to perform anchorperson detection. The system first prepares a variability matrix for I-vector and computes DNN parameters in the training phase using the speech from the target anchorperson and several imposter speakers. In the testing phase, the hybrid system determines the validity of the anchorperson with the testing utterance without any text information of the spoken materials. When the audio of collected video strings were recorded by different devices, the detection performance can be drastically degraded. To handle this issue, the present study also investigates robust feature schemes to handle the mismatch issue of recording devices. Traditionally, cepstral mean subtraction (CMS) [15], cepstral mean and variance normalization (CMVN) [16], and histogram equalization (HEQ) [17] approaches have been used to compensate for environmental mismatches. Subspace feature normalization (SFN) [18], a recently proposed robust front-end processing technique that produced satisfactory performance with high data compression capability, will be used to handle the mismatch issue.

This paper is organized as follows: Section 2 describes the speaker verification scenario, Section 3 presents the proposed hybrid I-vector and DNN system and SFN scheme, Section 4 reports the experimental setup and results, and Section 5 concludes the findings of this study.

II. SPEAKER VERIFICATION FOR ANCHORPERSON DETECTION

This section presents the structure of a talk show program and a speaker verification scenario that we used to perform the anchorperson detection. The speaker verification scenario is divided into two parts: I-vector extraction and modeling and classification.

A. Structure of Talk Show Program

Fig. 1 shows a structure of typical talk show programs [19]. The talk show generally has starting, content, and ending sessions. The content part has several stories, and each story contains A-shot (for anchorperson) and G-shot (for guest speakers). Based on the anchorperson detection function, one can determine the breakpoints of separate stories and therefore efficiently locate interesting topic segments of video from the entire video stream.

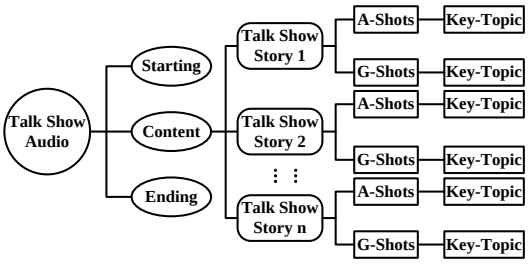

Fig. 1. *Structure of a typical talk show program. A-Shots: Anchorperson shots. G-Shots: Guests shots.*

B. Speaker Verification Scenario

Fig. 2 shows a speaker verification system. The feature extraction unit produces representative features from audio streams for both the training and testing data sets. The training features are used to estimate a model, which is then used to determine the validity of the claimed identity of the testing data. When applying this speaker verification system for anchorperson detection, the target model is estimated by the spoken data of the target anchorperson, and the background model is computed by using the data from imposter speakers, such as guests or other anchorpersons. The feature extraction and modeling and verification stages will be introduced in more detail in the following two sections.

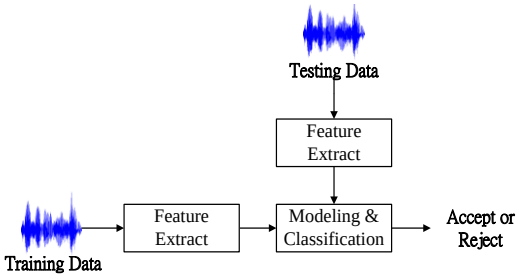

Fig. 2. *Speaker verification scenario.*

C. Feature Extraction

The goal of the feature extraction unit is to convert audio data into compact and representative feature vectors. In this study, the I-vector is used to characterize the speech segments. I-vector extraction contains the following three steps: First, each speech waveform is extracted into a series of acoustic features [e.g. Mel-frequency cepstral coefficient (MFCC)]. Next, a low rank total variability matrix, denoted as T in the following discussion, is estimated by the entire set of training data. Finally for a speech segment X, we can compute its I-vector by

$$M = m + Tv, \tag{1}$$

where M is the supervector representation, and v is a low dimensional vector, namely the I-vector, for that particular speech segment. Suppose that X contains L frames, $\{x_1, x_2, \ldots, x_L\}$, and an UBM Ω composed of C mixture components used to characterize the total variability matrix, T, the Baum-Welch algorithm computes the statistics required to estimate the I-vector by

$$N_c = \sum_{l=1}^{L} \mathrm{P}(c \mid x_l, \Omega); \tag{2}$$

$$F_c = \sum_{l=1}^{L} \mathrm{P}(c \mid x_l, \Omega) x_l; \tag{3}$$

$$\tilde{F}_c = \sum_{l=1}^{L} \mathrm{P}(c \mid x_l, \Omega)(x_l - m_c), \tag{4}$$

where c=1, 2…C, and m_c is the mean of UBM mixture component c. With the statistics in (2)-(4), I-vector for the given utterance X can be obtained using the following equation:

$$v = (I + T^t \Sigma^{-1} N(u) T)^{-1} T^t \Sigma^{-1} \tilde{F}(u), \tag{5}$$

where $N(u)$ is a diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are N_c (c=1,…C) identity matrices I, $\tilde{F}(u)$ is a supervector of dimension $CF \times 1$ obtained by concatenating all of the first-order Baum-Welch statistics $\tilde{F}_c$ for the given utterance, and Σ is a diagonal covariance matrix of dimension $CF \times CF$ estimated in the training stage. More details about I-vector approach can be found in [13, 14].

D. Modeling and Verification

The goal of this stage is to build models to characterize speakers and then verify the speaker's identity based on the testing utterance. Previous studies showed that SVM is a successful classifier for performing classification based on the extracted I-vectors [13, 14]. In this study, we use SVM as the baseline system. In the training phase, we prepare a set of training data $V = \{(v_1, y_1), (v_2, y_2), \ldots, (v_M, y_M)\}$, where $y_1, \ldots, y_M$ are the corresponding labels to training I-vectors $v_1, \ldots, v_M$.

In verification, the I-vector of a testing utterance v_{test} imports to SVM system and determine verify result. When a kernel function is used, SVM determines the classification result by

$$h(v) = \sum_{i=1}^{M} \alpha_i y_i k(v_{test}, v_i) + w_0, \tag{6}$$

where α_i and w_0 are parameters, and $k(\cdot,\cdot)$ is a kernel function.

III. HYBRID I-VECTOR AND DNN SYSTEM AND ROBUST FROND-END APPROACHES

This section introduces hybrid I-vector and DNN system and several well-known robust frond-end processes, and presents a recently proposed SFN approach. The goal of robust front-end processes is to reduce recording devices mismatches by mapping features to make them closer to each other. Hereafter the feature sequence is denoted as $C[n]$, where n is the frame index.

A. The Proposed Hybrid I-vector and DNN System

Neural network (NN) [12] originated in the 1950s. Scientists raised a "perceptron" neuron model to mimic the human brain's organization and operation. More recently, a series of DNN studies has attracted much attention for providing significant performance improvements on signal processing and pattern classification [20-22]. DNN is constructed by integrating multiple layers of NNs. Due to the deep structure formed by multiple layers, DNN provides significant improvements over NN when a sufficient amount of training data is available. Fig. 3 shows a DNN structure which comprises three types of layers: input, hidden, and output layers. DNN computes the posterior probability of input features and considers the category, which yields the highest posterior probability, to be the target class.

To ensure a stable performance, the parameters of DNN are generally initialized by the training data. Then, the back-propagation algorithm is applied to fine-tune the parameters. To avoid the over-fitting issue caused by insufficient training data, a dropout technique is often used for better performance.

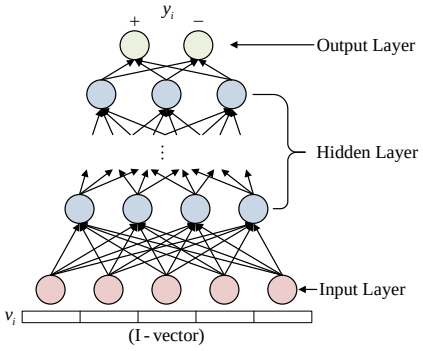

Fig. 3. *Structure of DNN.*

B. Conventional Feature Normalization Approaches

This section presents three well-known feature normalization approaches that are used in this study, including CMS, CMVN, and HEQ. CMS normalizes the mean of a feature by

$$\hat{C}[n] = C[n] - \mu, n = 1, \cdots, N, \quad (7)$$

where μ is the average of the sequence $C[n], n = 1, \cdots, N$, and N is the number of frames.

In addition to normalizing the mean, CMVN further normalizes the variance of the original feature sequence

$$\hat{C}[n] = \frac{C[n] - \mu}{\sigma}, n = 1, \cdots, N, \quad (8)$$

where σ is the standard deviation of the feature sequence.

HEQ uses a mapping function to convert the speech features to a referenced distribution by

$$\Phi(C[n]) = C_{tr}^{-1}(C_{ts}(C[n])), \quad (9)$$

where C_{ts} and C_{tr} are cumulative probability density and reference cumulative probability density functions, respectively, and $\Phi(\cdot)$ is the mapping function.

C. Subspace Feature Normalization

The SFN approach assumes that the noise components in the cepstral features are mainly located in high frequency bands [18]. Fig. 4 shows the overall procedure. First, SFN applies the discrete wavelet transform (DWT) to separate the original cepstral features into high frequency band (HFB) and low frequency band (LFB) parts. Next, SFN zeroes out the HFB part and normalizes the LFB part. Finally, an inverse DWT (IDWT) is applied to combine HFB and LFB parts to obtain normalized cepstral features.

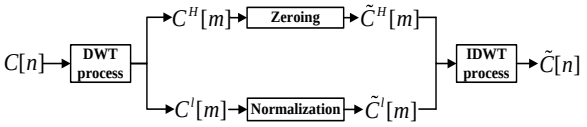

Fig. 4. *SFN flowchart.*

By applying DWT, the original $C[n]$ can be separated into HFB $C^H[m]$ and LFB $C^l[m]$ components

$$\{\tilde{C}^l[m], \tilde{C}^H[m]\} = DWT(C[n]), \quad (10)$$

where n=1…N, m=1…N/2, and N is the total number of frames. Then, SFN zeros out HFB components and applies a normalization algorithm on LFB components. Thus, we have

$$\begin{aligned} \tilde{C}^H[m] &= 0; \\ \tilde{C}^l[m] &= G(C^l[m]), \end{aligned} \quad (11)$$

where $\tilde{C}^l[m]$ and $\tilde{C}^H[m]$ are the processed LFB and HFB components, 0 represents a zero vector, and $G(\cdot)$ is a normalization function. Notably, the lengths of $\tilde{C}^l[m]$ and $\tilde{C}^H[m]$ are only half of the original cepstral feature stream because the down-sampling process is conducted in the DWT procedure. Finally, an IDWT is applied to transform the LFB and HFB components to form the normalized feature sequence:

$$\tilde{C}[n] = IDWT(\tilde{C}^l[m], \tilde{C}^H[m]). \quad (12)$$

From (12), since $\tilde{C}^H[m]$ is zeroed out, SFN achieves high compression for transmission data. Additionally, because the noise components in the HFB have been removed, SFN provides effective noise reduction in ASR (please refer to a previous study [18]). These two advantages make SFN particularly suitable for real-world mobile and distributed speech recognition applications.

IV. EXPERIMENT

A. Experimental Setup

In the experiment, we first compare the proposed hybrid I-vector and DNN system with the conventional hybrid I-vector and SVM system. Next, we test the effectiveness of using robust front-end processes for compensating for recording device mismatches. The SFN approach was also tested and compared with several conventional frond-end processes. To more accurately model the speech signals, we pre-process the training data by voice activity detector (VAD) to remove the silence signals.

a. Database

The experiment was designed to locate the segments pronounced by a target female anchorperson throughout the talk show. We collected 1921 utterances pronounced by 14 female speakers for training, and 118 utterances pronounced by the target anchorperson for the enrollment data. Test data included 300 target utterances and 540 guests (including both male and female speakers) utterances, where each utterance was around 3 seconds. Notably, the training, enrollment, and testing data were mutually exclusive.

In the experiment, the training data were first used to build a variability matrix for I-vector extraction. Then, the enrollment data and training data were used to estimate the parameters of DNN. In the verification, each testing segment was first converted to I-vector, which was then fed to DNN for testing verification.

b. Evaluation Method

Since the goal of this study is anchorperson detection, the first evaluation score is Accuracy, which is defined as

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}, \quad (13)$$

where tp , tn , fp ,and fn represent true positives, true negatives, false positives, and false negatives, respectively.

In addition to Accuracy, we also adopted the equal error rate (EER) and detection error tradeoff (DET) curves to compare performances. For speaker verification results, two types of detection error (false alarm rate and missed detection rate) may occur. The EER scores and DET curves can present the correlations of these two types of detection errors more directly.

c. Parameters setup

In this study, we used 39-dimensional feature vectors, including 13 static MFCCs and their first- and second- order dynamic features. The UBM-GMM contained 256 Gaussian components. The total variability matrix T contained 64 factors, thus producing a 64-dimensional I-vector for each speech segment. The DNN classifier included two hidden layers, each containing 150 neurons. The dropout technique was used with 20% inactive neurons. The SFN used the Haar functions [18] as the wavelet bases for DWT. The CMVN normalization was performed as $G(\cdot)$ in (11).

B. Experiment results

First, we compared I-vector integrated with SVM and with DNN, denoted as I-DNN and I-SVM, respectively. Table 1 shows the results of I-SVM and I-DNN in the second and third rows, respectively. In this set of experiments, the conventional MFCC features were used without performing frond-end processing. Meanwhile, Fig. 5 shows DET curves of I-SVM and I-DNN.

Table 1. *Accuracy and EER for SVM and DNN.*

	Accuracy (%)	EER (%)
I-SVM	78.14	21.85
I-DNN	82.82	19.63

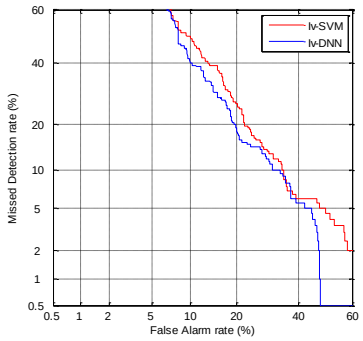

Fig. 5. *Comparison of SVM and DNN.*

From Table 1, I-DNN yields an Accuracy score of 82.82% and an EER score of 19.63%, which represent 4.68% and 2.22% absolute improvements compared to I-SVM, respectively. From Fig. 5, the two curves again show that I-DNN outperforms I-SVM, further confirming that DNN provides superior capability of speaker molding for this anchorperson detection task.

Next, we present the results of using different front-end processes in Table 2. The results of using CMS, CMVN, HEQ, and SFN with I-DNN are denoted as CMS+I-DNN, CMVN+I-DNN, HEQ+I-DNN, SFN+I-DNN, respectively. Meanwhile, Fig. 6 demonstrates the DET curves for these four results.

Table 2. *EER for various robust feature approaches.*

	Accuracy (%)	EER (%)
CMS+I-DNN	82.91	16.11
CMVN+I-DNN	85.09	17.50
HEQ+I-DNN	84.82	19.07
SFN+I-DNN	86.18	15.00

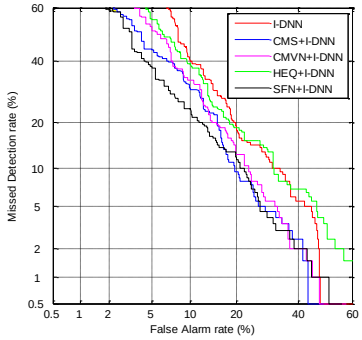

Fig. 6. *Various frond-end approaches.*

From the results, we obtain the following observations. First, a comparison of Table 1 and 2 reveals that all of the robust features can improve the Accuracy and EER scores, confirming the effectiveness of the front-end processing for compensating mismatches caused by recording devices. Next, from the results in Table 2 and Fig. 6, SFN+I-DNN achieves the best performance among the four approaches, confirming the superior capability of SFN to enhance performance robustness. Notably, since SFN only requires 50% amount of data, it is especially suitable to be applied in mobile devices for audio detection tasks.

V. CONCLUSION

In this paper, we proposed a robust speaker verification system using the hybrid I-vector and DNN system to perform anchorperson detection in video streams. We first compared the performances achieved by the integration of I-vector with SVM and I-vector with DNN. The experimental results demonstrated that I-vector with DNN provides 4.68% and 2.22% improvements of Accuracy and EER scores respectively, over the conventional I-vector with SVM. To handle the recording device mismatches, we incorporated robust front-end processes to the proposed hybrid system. The experimental results show that using robust front-end approaches can successfully overcome the mismatch issue. Finally, the results confirmed that SFN provides the best among related compensation approaches with 50% high data compression.

REFERENCES

[1] Haller, M., Kim, H. G. and Sikora, T., "Audiovisual anchorperson detection for topic-oriented navigation in broadcast news," in proc. ICME, pp. 1817-1820, 2006.
[2] Liu, Z. and Huang, Q., "Adaptive anchor detection using online trained audio/visual model," Electronic Imaging, pp. 156-167, 1999.
[3] Qi, W., Gu, L., Jiang, H. and Chen, X.-R., "Integrating visual, audio and text analysis for news video," in proc. ICIP, pp. 520-523, 2000.
[4] Gish, H. and Schmidt, M., "Text-independent speaker identification," IEEE Signal Processing Magazine, vol. 11, pp. 18-32, 1994.
[5] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.
[6] Reynolds, D. A., "An overview of automatic speaker recognition technology," in proc. ICASSP, pp. 4072-4075, 2002.
[7] Reynolds, D. A. and Rose, R., "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Transactions, Speech and Audio Processing, vol. 3, pp. 72-83, 1995.
[8] Svendsen, T. and Soong, F. K., "On the automatic segmentation of speech signals," in proc. ICASSP, pp. 77-80, 1987.
[9] Lee, C. H., Soong, F. K. and Juang, B. H., "A segment model based approach to speech recognition," in proc. ICASSP, pp. 501-541, 1988.
[10] Naik, J., Netsch, L. P. and Doddington, G. R., "Speaker verification over long distance telephone lines," in Proc. ICASSP, pp. 524-527, 1989.
[11] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E. and Torres-Carrasquillo, P. A., "Support vector machines for speaker and language recognition," Computer Speech & Language, vol. 20, pp. 210-229, 2006.
[12] Elman, J. L., "Learning and development in neural networks: the importance of starting small," Cognition, vol. 48, pp. 71-99, 1993.
[13] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., "Front-end factor analysis for speaker verification," IEEE Transactions, Audio, Speech, and Language Processing, vol. 19, pp. 788-798, 2011.
[14] Rao, W. and Mak, M. W., "Boosting the performance of I-vector based speaker verification via utterance partitioning," IEEE Transactions, Audio, Speech, and Language Processing, vol. 21, pp. 1012-1022, 2013.
[15] Viikki, O. and Laurila, K., "A recursive feature vector normalization approach for robust speech recognition in noise," in proc. ICASSP, pp. 733-736,1998.
[16] Tibrewala, S. and Hermansky, H., " Multiband and adaptation approaches to robust speech recognition," in Proc. Eurospeech, pp. 2619-2622, 1997.
[17] Hilger, F. and Ney, H., "Quantile based histogram equalization for noise robust large vocabulary speech recognition," IEEE Transactions, Audio, Speech and Language Processing, vol. 14, pp. 845-854, 2006.
[18] Wang, S.-S., Hung, J.-W. and Tsao, Y., "A study on cepstral sub-band normalization for robust ASR," in proc. ISCSLP, pp. 141-145, 2012.
[19] Gao, X. and Tang, X., "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," IEEE Transactions, Circuits and Systems for Video Technology, vol. 12, pp. 765-776, 2002.
[20] Larochelle, H., Benqio, Y., Louradour, J. and Lamblin, P., "Exploring strategies for training deep neural networks," Machine Learning, vol. 10, pp. 1-40, 2009.
[21] Bengio, Y., "Learning deep architectures for AI," Foundation and Trends in Machine Learning, vol. 2, pp. 1-127, 2009.
[22] Mohamed, A., Dahl, G. E. and Hinton, G. E., "Acoustic modeling using deep belief networks," IEEE Transactions, Audio, Speech, and Language Processing, vol. 20, pp. 14–22, 2013.