Face and Voice Recognition Using a Shared CoAtNet Framework: From Unimodal to Multimodal

Tuan-Duc Nguyen[1], Dinh-Dung Nguyen[1], Nguyen Anh Tuan[1], Cong-Hoang Diem[2], and Duc-Tho Mai[3,*]

[1] Swinburne Vietnam, FPT University, Hanoi, Vietnam
[2] Vietnam National University - International School, Hanoi, Vietnam
[3] Academy of Cryptography Techniques, Hanoi, Vietnam
`Corresponding mail: ductho-mai@actvn.edu.vn`

Abstract. Traditional authentication methods such as passwords, PINs, and fingerprint recognition face increasing challenges regarding security and usability. In response to these issues, this paper presents a deep learning-based biometric recognition system that integrates both facial and vocal modalities within a unified framework. We utilize CoAtNet, a hybrid convolution-attention architecture, as a shared backbone in a Siamese setup for face and voice recognition tasks. Experiments conducted on the Kaggle Face dataset, LFW, and LibriSpeech dev-clean demonstrate that the system performs exceptionally well independently: it achieved an accuracy of 97.78% for face recognition and 97.50% for voice recognition, with an Equal Error Rate (EER) of 2.44%. A version that combines both face and voice inputs exhibits lower accuracy due to the premature merging of the two data types, highlighting the importance of careful integration of different information types. Our findings validate the effectiveness of CoAtNet for unimodal biometric verification and indicate potential directions for enhancing cross-modal alignment in future research.

Keywords: Deep Learning, Face Recognition, Voice Recognition, Multimodal Recognition System

1 Introduction

Reliable identity verification remains a cornerstone of modern security systems, playing a critical role in domains ranging from personal device access to border control and financial transactions. Traditional authentication mechanisms, such as passwords, PINs, and hardware tokens, have long been the standard but suffer from limitations in both security and usability. Even biometric modalities such as fingerprint and iris recognition face practical issues such as spoofing, sensor degradation, and environmental variability. Recent survey studies [1], [2] have emphasized that unimodal authentication systems are often vulnerable to noise, pose variation, illumination changes, and adversarial attacks, especially in uncontrolled or real-world conditions.

To overcome these limitations, a range of learning-based and multimodal biometric systems has been explored. Deep learning approaches have substantially improved performance in face and voice recognition, with models like MagFace [3], CosFace [4], and Wav2Vec 2.0 [5] achieving state-of-the-art results on benchmark datasets. However, while these unimodal models excel in clean environments, they are still susceptible to failure when their respective inputs are corrupted. In response, researchers have proposed multimodal biometric systems that fuse face and voice modalities, leveraging their complementary characteristics to enhance robustness [6],[7]. Nevertheless, naive fusion strategies often suffer from modality misalignment, suboptimal shared representations, and inefficient architectural design.

In this study, we propose a unified Siamese architecture based on CoAtNet-0, a lightweight yet expressive convolution-attention hybrid backbone—for joint face and voice recognition. The system is trained on paired datasets: face images from a Kaggle face dataset and voice samples from the LibriSpeech `dev-clean` subset. Both modalities are encoded into 128-dimensional embeddings and trained using a contrastive loss to minimize intra-class distance while maximizing inter-class distance. We evaluate the model using accuracy ($\sim$ 97%), Equal Error Rate (2.44%), and parameter efficiency (22M), and compare our results against well-established unimodal and multimodal baselines. Experimental results demonstrate that our unimodal CoAtNet-0 models achieve competitive or superior performance with fewer parameters, while the multimodal fusion strategy reveals limitations that warrant further architectural refinement.

The remainder of this paper is organized as follows. Section 2 reviews related work in unimodal and multimodal biometric systems. Section 3 presents the proposed architecture and data processing pipeline. Section 4 describes the experimental setup and results evaluation. Finally, Section 5 concludes the study and outlines future directions for improving multimodal integration.

2 Related Works

Recent advances in deep learning have substantially improved the performance of biometric recognition systems across both facial and vocal modalities. In the domain of face recognition, several models such as FaceNet [8] and DeepFace [9] achieved 97–99% accuracy on standard benchmarks like LFW by leveraging convolutional neural networks trained with triplet loss or employing 3D alignment techniques, while the VGGFace2 benchmark [10] is a bit lower (92.0% accuracy, $\sim$1.0% EER). Building upon these foundations, subsequent models such as CosFace and MagFace introduced margin-based loss functions to enhance feature discriminability, achieving state-of-the-art performance with accuracy up to 99.83% and EERs below 0.05%. However, these models typically rely on deep and resource-intensive backbones like ResNet-100, resulting in over 40 million parameters.

In voice recognition, traditional systems based on deep neural networks (DNNs) and time-delay neural networks (TDNNs) [11] have gradually been supplanted by self-supervised approaches. Wav2Vec 2.0 [5], for example, utilizes Transformer-based architectures to achieve EERs as low as 2.5–3.0% on datasets such as VoxCeleb1. Despite their effectiveness, such models often require substantial computational resources, with parameter counts exceeding 90 million. Alternatively, lightweight baselines like VoicePrivacy [12] offer reduced complexity but suffer from significantly higher error rates (EER ∼11.81%).

To address the limitations of unimodal systems, an increasing body of research has turned to multimodal biometric authentication. Studies such as [6], [7] have demonstrated the effectiveness of combining face and voice modalities using conventional fusion strategies—ranging from feature concatenation with PCA and MFCC to deep models like AlexNet and VGGish. These approaches often report high accuracy (99%) on controlled datasets. Nonetheless, most of them operate at the score or decision level and do not align modalities in a shared feature space, limiting their generalizability.

More recent research has emphasized the need for modality-aware integration, advocating for joint embedding spaces or cross-modal attention mechanisms to enable coherent fusion of visual and auditory features. Building on this direction, our work introduces a unified Siamese architecture with a shared CoAtNet-0 backbone comprising only 22 million parameters. Unlike prior studies that rely on large and modality-specific encoders, our approach emphasizes compactness, shared representation, and consistent evaluation using accuracy, EER, and model size as core metrics.

3 The Proposed Methods

3.1 CoAtNet

Recent advancements in DL have highlighted the complementary strengths of convolutional neural networks (CNNs) and Transformer-based architectures in representation learning for both visual and auditory modalities. CNNs inherently capture local spatial patterns due to translation equivariance and local connectivity, enabling strong generalization, especially in low-data regimes. In contrast, Transformers utilize self-attention mechanisms to model long-range dependencies and global context, but they typically require larger datasets and heavier regularization due to the absence of spatial priors. To bridge these two paradigms, Dai *et al.* [13] proposed CoAtNet—a hybrid architecture that unifies convolution and attention mechanisms within a single framework. CoAtNet is built on two key principles: (1) depthwise convolution and self-attention can be integrated via relative attention, and (2) stacking convolutional blocks in the early stages, followed by Transformer blocks in the later stages, improves the trade-off between generalization and capacity.

Figure 1 illustrates the overall structure of CoAtNet, which consists of five processing stages:

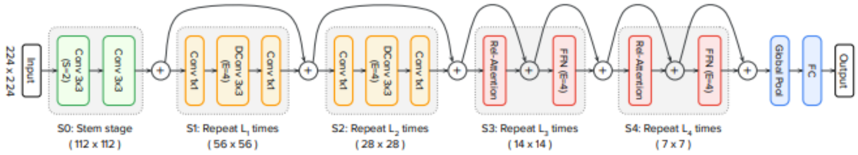

Fig. 1. CoAtNet Architecture

- **S0 (Stem stage):** Two 3×3 convolutional layers for initial downsampling and channel expansion.
- **S1–S2:** Repeated MBConv blocks (Mobile Inverted Bottleneck Convolution) for capturing local spatial features.
- **S3–S4:** Transformer stages incorporating relative self-attention and feed-forward networks (FFNs) for modeling long-range dependencies.
- **Final layers:** Global average pooling and a fully connected output layer.

We adopt CoAtNet-0, a compact version with approximately 22 million parameters, offering a good trade-off between efficiency and accuracy for biometric tasks.

3.2 Unimodal Biometric Recognition Using CoAtNet

We employ a shared CoAtNet-0 backbone for both face and voice recognition, following a Siamese learning structure.

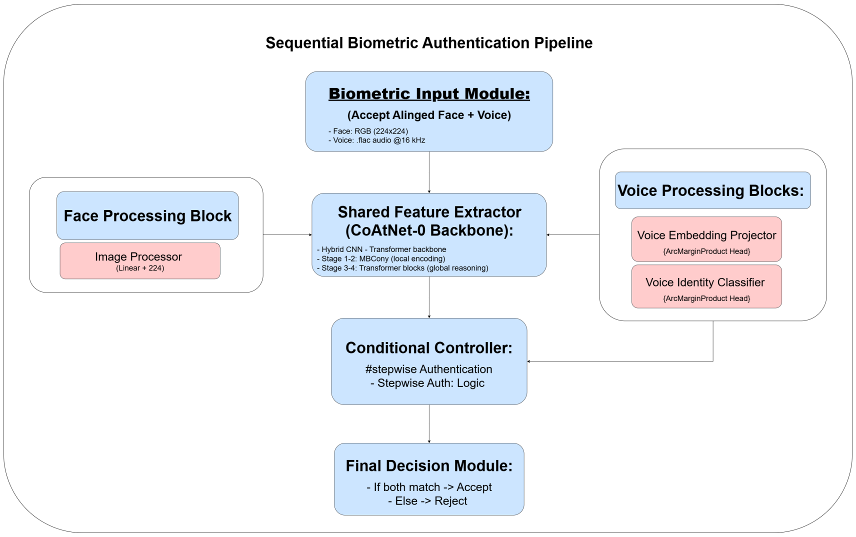
Sequential Biometric Authentication Pipeline
Biometric Input Module:
(Accept Alinged Face + Voice)
- Face: RGB (224x224)
- Voice: .flac audio @16 kHz
Face Processing Block
Image Processor
(Linear + 224)
Shared Feature Extractor (CoAtNet-0 Backbone):
- Hybrid CNN - Transformer backbone
- Stage 1-2: MBCony (local encoding)
- Stage 3-4: Transformer blocks (global reasoning)
Voice Processing Blocks:
Voice Embedding Projector
{ArcMarginProduct Head}
Voice Identity Classifier
{ArcMarginProduct Head}
Conditional Controller:
#stepwise Authentication
- Stepwise Auth: Logic
Final Decision Module:
- If both match -> Accept
- Else -> Reject

Fig. 2. Sequential Biometric Authentication Pipeline

Regarding the dataset, this paper utilizes two facial image datasets: (1) A curated dataset hosted on Kaggle, referred to as Face Recognition Dataset [14]; (2) The Labeled Faces in the Wild (LFW) dataset [15], a standard benchmark for face verification with over 13,000 images from 5,749 individuals, also available on Kaggle. To ensure consistent evaluation, we select 20 identities from each dataset with at least three images and apply an 80/20 train-test split. Each image is preprocessed using Albumentations: resized to 224 × 224, augmented with flipping, contrast enhancement, random rotation, and optional blur. Processed images are passed through the CoAtNet-0 backbone pretrained on ImageNet. Extracted features are projected to 128-dimensional embeddings via a linear projection and L2 normalization. Triplet loss is used with anchor-positive-negative triplets, and ArcMarginProduct is integrated to improve class separation. Verification is based on cosine distance.

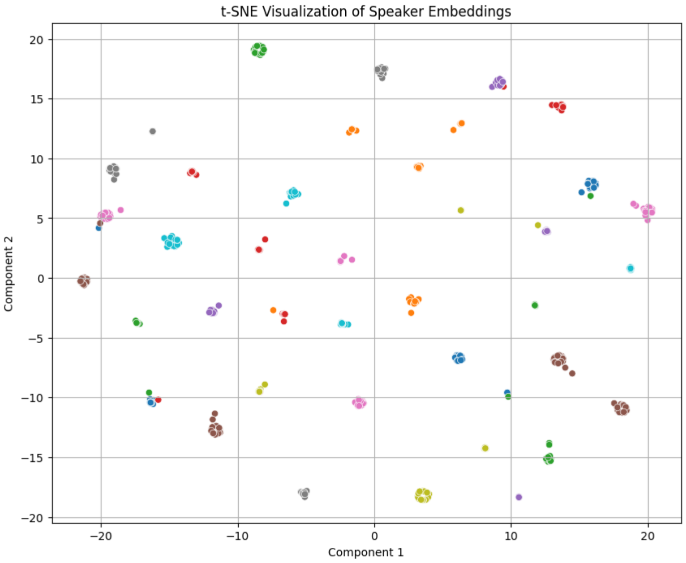

Fig. 3. t-SNE visualization of learned voice embeddings

Evaluation metrics include Accuracy, Precision, Recall, F1-score, False Accept Rate (FAR), False Reject Rate (FRR), and ROC-AUC. The voice recognition model uses the LibriSpeech `dev-clean` dataset [16]. Audio files in `.flac` format converted into Mel-Frequency Cepstral Coefficient (MFCC) spectrograms using 40 coefficients, a 25-ms frame size, and a 10-ms hop length. These MFCCs are normalized, stacked into three channels to simulate RGB, and resized to 224 × 224. Audio preprocessing is done using `librosa` and `torchaudio`. Each tensor is passed through the shared CoAtNet-0 backbone. Extracted features are projected to 128-dimensional embeddings using a linear layer and normalized with L2. A custom dataset builder is used to generate triplets for training with triplet loss. ArcMarginProduct is also employed to enhance speaker discrimination. Evaluation includes Accuracy, Equal Error Rate (EER), and ROC-AUC.

The t-SNE plot in Figure 3 shows clear speaker-specific clusters, confirming the discriminative quality of the voice encoder.

3.3 Multi-Modal

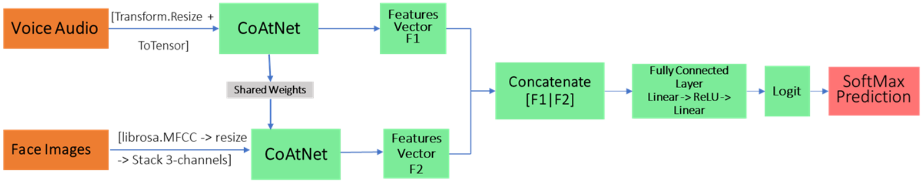

Fig. 4. Multimodal workflow: Parallel encoding and late fusion of face and voice embeddings.

A multimodal Siamese network that jointly embeds face and voice into a shared latent space to support identity verification is also proposed. As illustrated in Figure 4, the workflow accepts face images and voice spectrograms as parallel inputs. Each input is independently processed by a shared-weight CoAtNet-0 backbone, which extracts feature vectors F_1 (voice) and F_2 (face). These vectors are concatenated and passed through fully connected layers to generate the final prediction.

The face input is a resized RGB image (224×224), while the voice input is a 40-dimensional MFCC spectrogram mapped to a 3-channel pseudo-image. Both are processed through a CoAtNet-0 encoder that combines MBConv blocks and Transformer layers. Outputs from both streams are projected to 128-dimensional unit-normalized embeddings using a linear layer with L2-normalization.

To align face–voice embeddings, we employ CosineEmbeddingLoss:

$$\mathcal{L} = \begin{cases} 1 - \cos(z_1, z_2), & \text{if } y = +1 \\ \max(0, \cos(z_1, z_2) - m), & \text{if } y = -1 \end{cases} \tag{1}$$

where $m = 0.5$ is the margin, and y denotes identity match (+1) or mismatch (-1).

During inference, the system computes cosine similarity between embeddings for verification or retrieval. This design promotes modality-invariant representations and enables robust cross-modal matching in real-world settings.

4 Results and Discussion

4.1 Experimental Procedures

This section outlines the experimental setup, training configurations, evaluation metrics, and data collection used to benchmark the proposed models across unimodal (face, voice) and multimodal biometric recognition tasks. Experiments were conducted on a high-performance GPU environment, and models were trained using contrastive objectives in a Siamese architecture to assess identity verification accuracy, robustness, and generalization under real-world conditions.

Table 1. Computational Platform and Training Configuration

Component	**Configuration**
GPU	NVIDIA Tesla T4 (16GB VRAM)
Framework	PyTorch 2.0, torchaudio, torchvision
Input Size	224 × 224 × 3
Audio Representation	40-dimensional MFCCs (25ms window, 10ms stride) resized to 224 × 224 × 3
Image Representation	RGB face images resized to 224 × 224 × 3
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	16 pairs per batch (8 positive + 8 negative)
Epochs	20
Loss Function	CosineEmbeddingLoss (margin = 0.5)
Embedding Dimension	128

For each training iteration, pairs of samples (either face–face, voice–voice, or face–voice in multimodal setup) were passed through a shared CoAtNet-0 backbone. The model was trained to minimize cosine distance for matching pairs and maximize it for non-matching ones. Data augmentation techniques such as random horizontal flipping (for images) and time masking (for audio) were applied to enhance generalization.

Model performance was evaluated using accuracy, Equal Error Rate (EER), and ROC-AUC metrics on hold-out test sets. All experiments were repeated using the same random seed to ensure consistency. Metrics were logged per epoch and used to compare unimodal versus multimodal strategies under the same configuration.

4.2 Results and Analytics

The unimodal CoAtNet-0 models demonstrate strong performance in both face and voice recognition tasks with a relatively lightweight architecture. Our face recognition model achieved 97.78% accuracy and an EER of 0.4%, outperforming the VGGFace2 benchmark [10] (92.0% accuracy, ∼1.0% EER) while using fewer

Table 2. Comparison table

Model	Modality	Backbone	Loss Function	Embedding	Params (M)	Accuracy (%)	EER (%)
Our Face (CoAtNet-0)	**Face**	**CoAtNet-0**	**Triplet + ArcMargin**	**128D**	**22**	**97.78**	**0.4**
VGGFace2 [10]	Face	ResNet-50	Softmax + Cosine	512D	∼26	92.0 (IJB-B)	∼1.0
CosFace [4]	Face	ResNet-100	Cosine Margin	512D	∼42	99.73 (LFW)	∼1.0
MagFace [3]	Face	ResNet-100	MagFace Loss	512D	∼42	99.83 (LFW)	0.037
Our Voice (CoAtNet-0)	**Voice**	**CoAtNet-0**	**Triplet + ArcMargin**	**128D**	**22**	**97.50**	**2.44**
Wav2Vec 2.0 [5]	Voice	Transformer	CTC + LM / Cosine	768–1024D	∼94	96.5	2.7
VoicePrivacy [12]	Voice	CNN + x-vector	CE / PLDA	—	∼12	85.0	11.81

parameters (22M vs. ∼26M). While CosFace [4] and MagFace [3] report higher accuracy on the LFW dataset (99.73% and 99.83%, respectively), these models rely on significantly deeper backbones (ResNet-100) and over 40M parameters, highlighting the efficiency–performance trade-off of our approach.

In the voice modality, our model trained on LibriSpeech `dev-clean` achieved 97.50% accuracy and a low EER of 2.44%. This result is comparable to Wav2Vec 2.0, which achieves 96.5% accuracy and around 2.7% EER, but with more than four times the parameter count (94M vs. 22M). Moreover, our model significantly outperforms the lightweight VoicePrivacy baseline, which reports only 85.0% accuracy and 11.81% EER.

Despite these strong results in the unimodal setting, our multimodal model, which combines face and voice embeddings via early fusion, yields notably lower performance: 62.59% accuracy and 68.96% EER. This drop indicates a misalignment between modalities in the joint embedding space, consistent with prior findings that naive fusion strategies often fail to exploit cross-modal complementarities effectively. Future work should explore late fusion or attention-based mechanisms to better integrate heterogeneous features.

5 Conclusion

In this study, we suggested a system that uses deep learning to verify identity by combining face and voice data in a special network design called a Siamese network, which is supported by CoAtNet-0. The unimodal systems demonstrated strong performance, achieving 97.78% accuracy for face recognition and 97.50% accuracy for voice verification, with an equal error rate (EER) of 2.44%. The results show that CoAtNet-0 is very good at identifying important features for both face and voice data. However, the multimodal variant that utilized early embedding fusion exhibited a significantly lower performance, with an accuracy of only 62.59%. This indicates that simplistic fusion strategies can result in misaligned feature spaces, particularly under varying input conditions such as noise, lighting variations, and discrepancies in hardware quality. To tackle these issues, future research will explore better ways to combine data, such as (1) attention mechanisms that focus on the quality of the input, (2) balancing loss to make sure both types of data are equally important during training, and (3) methods to share and align knowledge between face and voice data. These enhancements aim to improve the robustness, adaptability, and overall performance of multimodal biometric systems in real-world environments.

References

[1] L. Cherrat, M. Kardouchi, and A. Ouahabi, "A survey on biometric authentication systems based on face and voice," *IEEE Access*, vol. 8, pp. 210 192–210 211, 2020. DOI: 10.1109/ACCESS.2020.3039739.

[2] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2019. DOI: 10.1145/3038924.

[3] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, *Magface: A universal representation for face recognition and quality assessment*, 2021. arXiv: 2103.06627.

[4] H. Wang, Y. Wang, Z. Zhou, *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274. DOI: 10.1109/CVPR.2018.00552.

[5] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020.

[6] F. Abbaas and G. Serpen, *Evaluation of biometric user authentication using an ensemble classifier with face and voice recognition*, 2020. arXiv: 2006.00548 [cs.CR].

[7] M. Saleh and I. Jouny, "Multimodal person identification through the fusion of face and voice biometrics," in *2022 17th Annual System of Systems Engineering Conference (SOSE)*, 2022, pp. 164–169. DOI: 10.1109/SOSE55472.2022.9812670.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, " FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220.

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, " VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recogni-*

tion (FG 2018), Los Alamitos, CA, USA: IEEE Computer Society, May 2018, pp. 67–74. DOI: 10.1109/FG.2018.00020.

[11] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.

[12] N. Tomashenko, X. Wang, X. Miao, *et al.*, *The voiceprivacy 2022 challenge evaluation plan*, 2022. arXiv: 2203.12468.

[13] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21, Red Hook, NY, USA: Curran Associates Inc., 2021, ISBN: 9781713845393.

[14] *Face Recognition Dataset*, en. [Online]. Available: https://www.kaggle.com/datasets/vasukipatel/face-recognition-dataset (visited on 07/21/2025).

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, IEEE, 2015, pp. 5206–5210.