Fusion of Classical Digital Signal Processing and Deep Learning Methods (FTCAPPS)

Gomez, Angel; Sánchez, Victoria E.; Peinado, Antonio; Martín-Doñas, Juan M.; Gómez-Alanís, Alejandro; Villegas-Morcillo, Amelia; Rosello, Eros; Chica, Manuel; Garcia, Celia; Espejo, Ivan Lopez

Published in:
Proc. IberSPEECH 2022

DOI (link to publication from Publisher):
10.21437/IberSPEECH.2022-48

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Gomez, A., Sánchez, V. E., Peinado, A., Martín-Doñas, J. M., Gómez-Alanís, A., Villegas-Morcillo, A., Rosello, E., Chica, M., Garcia, C., & Espejo, I. L. (2022). Fusion of Classical Digital Signal Processing and Deep Learning Methods (FTCAPPS). In *Proc. IberSPEECH 2022* (pp. 237-240) https://doi.org/10.21437/IberSPEECH.2022-48

Fusion of Classical Digital Signal Processing and Deep Learning methods (FTCAPPS)

Angel M. Gomez[1], *Victoria E. Sánchez*[1], *Antonio M. Peinado*[1], *Juan M. Martín-Doñas*[2],
Alejandro Gómez-Alanis[3], *Amelia Villegas-Morcillo* [1], *Eros Rosello*[1], *Manuel Chica*[1],
Celia Garcia[1], *Iván López-Espejo*[4]

[1] Universidad de Granada (Spain),
[2] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia-San Sebastián (Spain),
[3] Amazon Alexa (Germany),
[4] Aalborg University (Denmark)

{amgg,victoria,amp,ameliavm, erosrosello, manuelc, cgr14}@ugr.es,
jmmartin@vicomtech.org, agomezalanis@amazon.de, ivl@es.aau.dk

Abstract

The use of deep learning approaches in Signal Processing is finally showing a trend towards a rational use. After an effervescent period where research activity seemed to focus on seeking old problems to apply solutions entirely based on neural networks, we have reached a more mature stage where integrative approaches are on the rise. These approaches gather the best from each paradigm: on the one hand, the knowledge and elegance of classical signal processing and, on the other, the great ability to model and learn from data which is inherent to deep learning methods. In this project we aim towards a new signal processing paradigm where classical and deep learning techniques not only collaborate, but fuse themselves. In particular, we focus on two objectives: 1) the development of deep learning architectures based on or inspired by signal processing schemes, and 2) the improvement of current deep learning training methods by means of classical techniques and algorithms, particularly, by exploiting the knowledge legacy they treasure. These innovations will be applied to two socially and scientifically relevant topics in which our research group has been working for years. The first one is the enhancement of speech signal acquired under acoustic adverse conditions (e.g., noise, reverberation, other speakers, ...). The second one is the development of anti-fraud measures for biometric voice authentication, in which banking corporations and other large companies are strongly interested.

Index Terms: Machine Learning, Deep Neural Networks, Speech Enhancement, Multichannel speech processing, Voice Anti-spoofing

1. Introduction

In the last decade we have witnessed a radical change in the classical framework where, for the second half of the last century, signal processing founded itself as a new discipline. The need to overcome the limitations associated with the classical assumptions of linearity and stationarity, as well as the extensive use of second-order statistics, has paved the way for the irruption of machine learning techniques and, in particular, Deep Neural Networks (DNN) [1]. Although they were previously known and applied, neural networks have found fertile ground in the last decade thanks to the advances, both algorithmic and from hardware, needed to handle the huge amounts of data they require.

Although, at first, there was a proliferation of solutions in which DNNs were conceived as black box units that performed a particular task from end-to-end, it was soon realized that these approaches wasted the accumulated in-domain knowledge we already had [2]. The efforts of the signal processing research community to take advantage of this knowledge within deep learning approaches have been remarkable in recent years [3, 4, 5, 6]. Our research team stakes on this integrative approach aiming to take a step further towards a new paradigm in which deep learning techniques are *seamlessly* incorporated into the existing signal processing theoretical body of knowledge.

To this end, we propose two lines of work: i) the design of new network architectures inspired by classical signal processing techniques, and ii) the development of new training methods that take advantage of available in-domain knowledge.

In the first case we look for new architectures, or the modification of existing ones, at the structural level, taking advantage of the available knowledge about the signals involved. Furthermore, we aim at the integration of advanced algorithms (e.g. adaptive filtering, Kalman filtering, etc.) at the layer (or cell) level, considering these as signal processing operators (in the same way that a convolutional layer can be understood as a filter-bank or a pooling as a decimation) [7]. Embedding these algorithms in a layer, or cell, would avoid the need for a heuristic adjustment of the algorithm parameters. These can directly be learned from data, while allowing us to non-linearly process signals, as the very DNNs naturally do.

In the second line of work we try to imbue DNNs with in-domain knowledge by means of the training procedure. The easiest way to do so is through the loss function used as training criterion. Thus, we propose both to adapt known classical quality or performance metrics that distill prior knowledge available on the signal, and to generate others adapted to the problem under consideration. Also, we propose the development of new data generation techniques (data augmentation) that make training possible or that, conveniently used during training, reduce the risk of overfitting. These generation techniques are based on simulation and, therefore, require knowledge of the problem and the associated classical signal processing techniques.

The resulting developments will be applied on two topics of current interest in which our research team has been working for years: multichannel speech enhancement and detection of

10.21437/IberSPEECH.2022-48

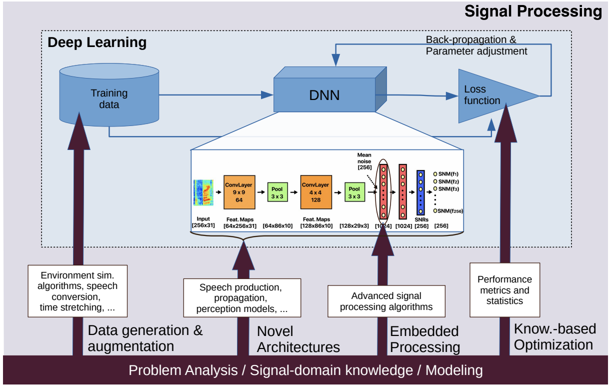
Signal Processing
Deep Learning
Training data
DNN
Loss function
Back-propagation & Parameter adjustment
Input [256x31]
ConvLayer 9 x 9 64
Feat. Maps [64x256x31]
Pool 3 x 3
[64x86x10]
ConvLayer 4 x 4 128
Feat. Maps [128x86x10]
Pool 3 x 3
[128x29x3]
Mean noise [256]
[1024] [1024]
SNRs [256]
SNM(f1) SNM(f2) SNM(f3) SNM(f256)
[256]
Environment sim. algorithms, speech conversion, time stretching, ...
Speech production, propagation, perception models, ...
Advanced signal processing algorithms
Performance metrics and statistics
Data generation & augmentation
Novel Architectures
Embedded Processing
Know.-based Optimization
Problem Analysis / Signal-domain knowledge / Modeling

Figure 1: *Key elements where classical knowledge-based signal processing methods and deep learning approaches can potentially be fused.*

malicious impersonation attacks in speech-based biometric systems. However, other possible applications in which our team also has experience, such as speech, image or video restoration, as well as proteomic signal processing, are not ruled out.

2. Project goals

As mentioned above, over the last decade deep neural networks have unquestionably demonstrated their suitability as a modeling tool in signal processing problems. In a precursor project [8] we showed that deep learning techniques could be successfully integrated with classical signal processing methods, combining the best of both approaches. In this project we take a step further and pursue a *seamless* integration in which deep learning models assimilate in an adequate way the knowledge accumulated over decades and crystallized in classical methods. In this sense, we not only pursue the integration of DNNs as a block within the processing pipeline, but also the fusion of elements across both approaches. Potential key elements which can enable this seamless integration are depicted in Figure 1. This way we intend to advance towards a new paradigm where classical signal processing techniques and deep learning are merged.

To achieve this, we primarily focus on the speech processing domain, in which our research team has extensive experience. In particular, we will consider two applications of particular interest at present. The first one is the enhancement and improvement of speech signal, a regression problem which has additional side applications such as speech or speaker recognition, and speech transmission. In addition, the widespread use of devices integrating multiple microphones (such as smartphones, tablets, home assistants and smartTVs) has put the spotlight on multichannel processing. The second one is the detection of malicious attacks against speaker identification systems, or anti-spoofing methods. This is a classification problem where we must discriminate between genuine or spoofed voice (prerecorded signal replayed or synthetic/converted speech). The goal is to avoid fraud in voice-based biometric systems, providing a solution to the reinforced authentication requirements demanded by today's society.

Development of new network architectures inspired by classical signal processing techniques

Our aim is to develop new architectures that replicate the macromodeling of the signal by means of prior knowledge about it. In the context of speech processing, we can take advantage of classical schemes of speech production, signal propagation and auditory perception models to establish the topology and processing units to be used in DNNs for speech enhancement and antispoofing applications [3, 9, 10]. Moreover, signal processing methods can even be embedded within the network layers. Indeed, certain layers already replicate well-known classical processing methods, as mentioned in the previous section. Our goal here is to translate more advanced methods from classical signal processing (adaptive filtering, Kalman filters, etc.) as network layers (or cells) embedded in the very DNN layers, so that we can benefit from nonlinear processing, while parameters can be adjusted during learning and relationships across data exploited.

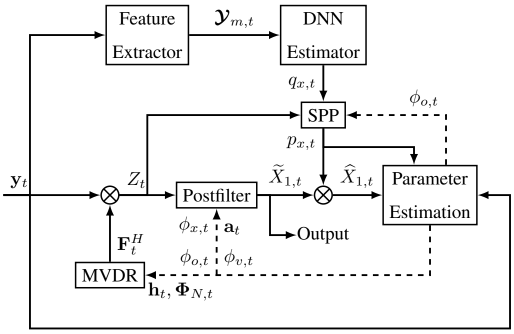

Figure 2: *Block diagram of the proposed Recursive EM algorithm for multichannel speech enhancement. The dashed lines mean the feedback due to the M-step of the algorithm.*

Novel knowledge-based training methods

We also explore the use of classical processing techniques for the development of generative and data-augmentation methods, which can alleviate the enormous input data needs that DNNs require for generalization. In addition to simple noise addition, other more advanced classical methods such as acoustic environment simulation, speech synthesis, speech conversion, time-stretching, bandwidth extension, etc. can be used [11]. Moreover, the development of loss functions for DNN training derived from classical speech quality and/or intelligibility metrics are of special interest [6], as well as those learning schemes that can generate enriched embeddings to facilitate classification and detection [12, 13].

3. Ongoing work

As a result of this project, we have fused architectures combining recurrent and convolutional neural networks for speech/noisy mask estimation with statistical signal processing algorithms. Thus, an architecture based on extended Kalman filtering has been proposed for two-channel speech enhancement [14], while an expectation-maximization (EM) algorithm combining neural network speech prediction with Kalman filtering [15] has been developed for multichannel enhancement in general (Figure 2).

For robust speaker recognition and verification, an antifraud system has been developed that combines convolutional and recurrent networks that integrates a single-channel noise estimator, providing significant improvements over the state of the art [16]. In addition, we have proposed a new adversarial transformation network based on GANs which is capable of generating adversarial spoofing attacks that, used during train-

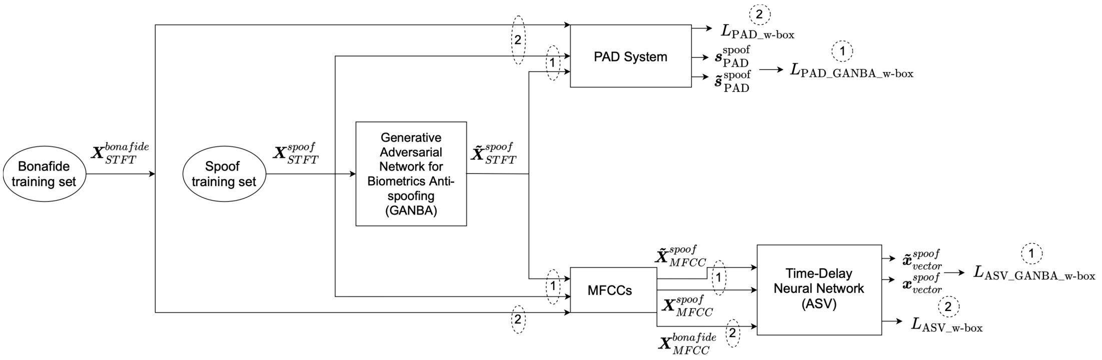
Bonafide training set
$\boldsymbol{X}^{bonafide}_{STFT}$
Spoof training set
$\boldsymbol{X}^{spoof}_{STFT}$
Generative Adversarial Network for Biometrics Anti-spoofing (GANBA)
$\tilde{\boldsymbol{X}}^{spoof}_{STFT}$
2
1
PAD System
$L_{\text{PAD\_w-box}}$
$\boldsymbol{s}^{\text{spoof}}_{\text{PAD}}$
$\tilde{\boldsymbol{s}}^{\text{spoof}}_{\text{PAD}}$
$L_{\text{PAD\_GANBA\_w-box}}$
MFCCs
$\tilde{\boldsymbol{X}}^{spoof}_{MFCC}$
$\boldsymbol{X}^{spoof}_{MFCC}$
$\boldsymbol{X}^{bonafide}_{MFCC}$
Time-Delay Neural Network (ASV)
$\tilde{\boldsymbol{x}}^{spoof}_{vector}$
$\boldsymbol{x}^{spoof}_{vector}$
$L_{\text{ASV\_GANBA\_w-box}}$
$L_{\text{ASV\_w-box}}$

Figure 3: *Generative adversarial network for biometric anti-spoofing (GANBA) framework for white-box scenarios. Step 1: generator-only training (ASV and PAD parameters frozen). Step 2: discriminator (ASV + PAD) training. Encircled outputs corresponding to classical cross-entropy loss function.*

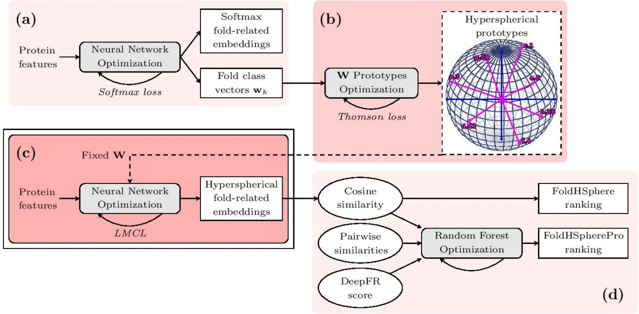
(a)
Protein features
Neural Network Optimization
Softmax loss
Softmax fold-related embeddings
Fold class vectors $\mathbf{w}_k$
(b)
W Prototypes Optimization
Thomson loss
Hyperspherical prototypes
(c)
Fixed W
Protein features
Neural Network Optimization
LMCL
Hyperspherical fold-related embeddings
Cosine similarity
Pairwise similarities
DeepFR score
Random Forest Optimization
FoldHSphere ranking
FoldHSpherePro ranking
(d)

Figure 4: *Overview of the FoldHSphere approach for protein fold recognition. In the first stage a we train a neural network model using the softmax cross-entropy as loss function. We then optimize the position of the classes by our Thomson-based loss, so that they are maximally separated in the angular space. The resulting hyperspherical prototypes are used as a fixed non-trainable classification matrix.*

ing, improve the state-of-the-art systems [17] (Figure 3).

We have also investigated the use of the Large Margin Cosine Loss (LMCL) function for extraction of prototype class vectors in deep neural network training, as well as its improvement by means of the adaptation of quasi-optimal solutions to the Thomson problem in order to achieve more representative embeddings (Figure 4). We have also applied this approach to other topics as proteomic signal processing, where it was very successful for protein folding type classification [18].

Finally, techniques developed by our team have been adapted for participation in the DiCOVA 2021 Challenge, whose objective was the detection of COVID19 from audio signal (cough recordings in *track 1*). The proposed system integrated classical techniques and neural networks for signal preprocessing and cough segment detection, as well as a score fusion system between multiple classifiers [19].

4. Conclusions and future work

In this paper we have presented the project FTCAPPS, which will be executed in the period from July 2020 till June 2023. The project involves researchers from the University of Granada in collaboration with expert researchers from other countries and from the industry.

Currently the project is providing novel approaches which exploit this fused approach of paradigms, as shown in the previous section. We aim at going further in this line and achieve groundbreaking developments which seamlessly integrate DNN as part of speech signal processing. Updated information about this project can be found at http://sigmat.ugr.es/proyectos/ftcapps.

5. Acknowledgements

This work has been supported by the project PID2019-104206GB-I00 funded by MCIN/ AEI /10.13039/ 501100011033.

6. References

[1] S. Haykin, “Neural Networks Expand SP’s Horizons”, *Signal Processing Magazine*, pp. 24 - 49, 1996.

[2] E. Vincent, “Is audio signal processing still useful in the era of machine learning?,” *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), pp. 7-7, 2015.

[3] Y. Xu, J. Du, L.R. Dai, C.H. Lee, “An experimental study on speech enhancement based on deep neural networks”, *IEEE Signal Process. Letters* vol. 21, no. 1, pp. 65-68, Jan. 2014.

[4] K. Tan, J. Chen, D. Wang, “Gated residual networks with dilated convolutions for monaural speech enhancement”, *IEEE Trans. on Audio, Speech, and Language Processing* vol. 27, no. 1, pp. 189-198, Jan. 2019.

[5] D. Wang, J. Chen, “Supervised speech separation based on deep learning: An overview”, *IEEE Trans. on Audio, Speech, and Language Processing*, 2018.

[6] J.M. Martín-Doñas, A.M. Gomez, J.A. Gonzalez, A.M. Peinado, “A Deep Learning Loss Function based on the Perceptual Evaluation of the Speech Quality”, *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680-1684, Nov. 2018.

[7] A.K. Maier, C. Syben, et al. “Learning with known operators reduces maximum error bounds”, *Nature machine intelligence*, 1(8), 2019.

[8] A.M. Gomez, V. Sanchez, et al. “Nuevas aproximaciones conexionistas integradas para procesado de señal”, Ministerio de Economía y Competitividad: Programa estatal de fomento de la investigación científica y técnica, 2016-2019, (TEC2016-80141-P).

[9] D. Wang, G.J. Brown, “Computational Auditory Scene Analysis: Principles, Algorithms, and Applications”, Hoboken, NJ, USA: Wiley, 2006.

[10] Y. Qian, N. Chen, K. Yu, “Deep Features for automatic spoofing detection”, *Speech Communication*, 85, 2016.

[11] F. Ma, L. Chai, J. Du, D. Liu, Z. Ye , C.-H. Lee, "Acoustic Model Ensembling Using Effective Data Augmentation for CHiME-5 Challenge", in Proc. *InterSpeech* 2019.

[12] S. Novoselov, V. Shchemelinin, A. Shulipa, A. Kozlov, I. Kremnev, "Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition", in Proc. *InterSpeech* 2018.

[13] R. Li, M. Zhao, Z. Li, L. Li, Q. Hong, "Anti-Spoofing Speaker Verification System with Multi-Feature Integration and Multi-Task Learning", in Proc. *InterSpeech* 2019.

[14] J.M. Martín-Doñas, A.M. Peinado, I. López-Espejo, A.M. Gómez, "Dual-channel eKF-RTF framework for speech enhancement with DNN-based speech presence estimation", in proceedings of *IBERSPEECH 2021*, pp. 31-35, 24-25 March 2021.

[15] J.M. Martín-Doñas, J. Jensen, Z.-H. Tan, A. M. Gómez, A. M. Peinado, "Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3080-3094, 2020.

[16] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. Pavankumar Dubagunta, A. M. Peinado, M. Magimai-Doss, "On Joint Optimization of Automatic Speaker Verification and Anti-spoofing in the Embedding Space", *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579-1593, 2021.

[17] A. Gomez-Alanis, J. A. Gonzalez-Lopez, A. M. Peinado, "GANBA: Generative Adversarial Network for Biometric Anti-spoofing", *MDPI Applied Sciences*, vol. 12, no. 3, pp. 1454, 2022.

[18] A. Villegas-Morcillo, V. Sanchez, A. M. Gomez, "FoldHSphere: Deep Hyperspherical Embeddings for Protein Fold Recognition", *BMC Bioinformatics*, 22(490), pp. 1-21. 2021.

[19] D. Madhu R. Kamble, J. A. Gonzalez-Lopez, T. Grau, J. M. Espin, L. Cascioli, Y. Huang, A. Gomez-Alanis, J. Patino, R. Font, A. M. Peinado, A. M. Gomez, N. Evans, M. A. Zuluaga, M. Todisco, "PANACEA cough sound-based diagnosis of COVID-19 for the DiCOVA 2021 Challenge", in proceedings of *INTERSPEECH 2021*, pp. 906-910, 2021.