Transfer Learning from Audio Domains a valuable tool for Structural Health Monitoring

Eleonora M. Tronci[1], Homayoon Beigi[2], Maria Q. Feng[1], Raimondo Betti[1]
[1] Department of Civil Engineering and Engineering Mechanics
Columbia University, New York, NY 10027
[2] Recognition Technologies, Inc.

ABSTRACT

Today, the application of artificial neural network tools to define models that mimic the dynamic behavior of structural systems is a wide-spread approach. A fundamental issue in developing these strategies for damage assessment in civil structures is represented by the unbalanced nature of the available databases, which commonly contain plenty of data coming from the structure under healthy operational conditions and very few samples from the system in unhealthy conditions since the structure would have failed by that time. Consequently, the learning task, carried on with standard deep learning approaches, becomes case-dependent and tends to be specialized for a particular structure and for a very limited number of damage scenarios.

This work presents a framework for damage classification in structural systems intended to overcome such limitations. In this methodology, the model is trained to gain knowledge in the learning task from a rich acoustic dataset (source domain), acquiring higher-level features characterizing vibration traits from a rich acoustic dataset. This knowledge is then transferred to a target domain, with much less training data, such as a structural system, in order to classify its structural condition.

The framework starts with constructing a Time-Delay Neural Network (TDNN) structure, trained on the VoxCeleb dataset, in the speech domain. The input of the network consists of Cepstral and pitch features extracted from the audio records. Higher-level features, the x-vectors, speaker embeddings, capturing neural outputs of the network's intermediate layers, are derived and then used to train a Probabilistic Linear Discriminant analysis (PLDA) model to provide a probabilistic discriminant model for speaker comparison. These features collect generic information regarding the source domain and characterize a classification process based on the frequency content of signals, which is not strictly dependent on the original acoustic domain. Because of the non-case-dependent nature of the x-vector embeddings (features), they can be used to train an alternative PLDA model to address a damage classification task, considering vibration measurements coming from a different system, a structural one which represents the target domain. The simulated data from the 12 degrees of freedom benchmark shear-building structure provided by the IASC-ASCE Structural Health Monitoring Group are studied to verify the proposed framework's effectiveness.

Keywords: Transfer Learning, Structural Health Monitoring, Mel-Frequency Cepstral Coefficients, Time-Delay Neural Network, x-vector features.

INTRODUCTION

The aim of vibration-based Structural Health Monitoring (SHM) methods is to implement a strategy to correctly detect damage by assessing changes in the identified vibration response of civil structures [1]. Identifying structural damage scenarios through information extracted from the structure's dynamic response can be addressed from different perspectives. An efficient way to do so is to apply Pattern Recognition and Machine Learning strategies (PRML) [1] that focus on detecting meaningful patterns in the features, defined as Damage Sensitive Features (DSFs), representing the conditions of the structural system.

A fundamental issue in the development of data-based PRML strategies for damage assessment is represented by the fact that the datasets available for mechanical and civil applications are quite limited in number and the amount of information provided. Unfortunately, when the aim is to classify between healthy and damaged conditions in a structure, it is extremely rare to have enough data for training the learning algorithm to recognize the unhealthy (or damaged) condition. This lack of information can cause ill-conditioning in training classification and prediction models built using machine learning and deep learning algorithms. Namely, the dataset does not contain sufficient information from all the different classes that need to be identified. Consequently, for mechanical and civil applications, the learning task becomes case-dependent and tends to be specialized for a particular structure and a very limited number of damage scenarios.

To overcome such a limitation imposed by the insufficiency of exhaustive failure datasets for civil and mechanical systems, we propose to take advantage of the richness and knowledge in the learning task gained from a source domain, with extensive and exhaustive datasets, and to transfer that same knowledge to a target domain, with much less information. This operation goes under the name of "transfer learning" and is emerging as a fundamentally new way of thinking in this data-driven and AI-centric era to move data around to enhance knowledge extraction effectively. Transfer Learning (TL) consists of transferring knowledge from one domain to another to take advantage of that knowledge for a different but related task or domain. The reader is referred to [2, 3, 4] for an overview of the different transfer learning methods. In this work, a transfer learning method is implemented to gain knowledge from automatic speech recognition, adapting the learning task to various classification tasks in structural domains or datasets. One of the advantages of deep learning is acquiring a hierarchy of feature representations from low-level features to more abstract higher-level features ones [5, 4]. Therefore, pre-training [6, 7, 8] on a rich domain to explicitly learn intermediate-level features in the neural network can be useful for several different tasks. The intermediate layers in a neural network's architecture trained on speech data appear to be not specific to any particular task, while the higher layers are task-specific [9]. Supervised training using out-of-domain data is also a form of pre-training, and it has been used to learn multilingual bottleneck features in [10, 11].

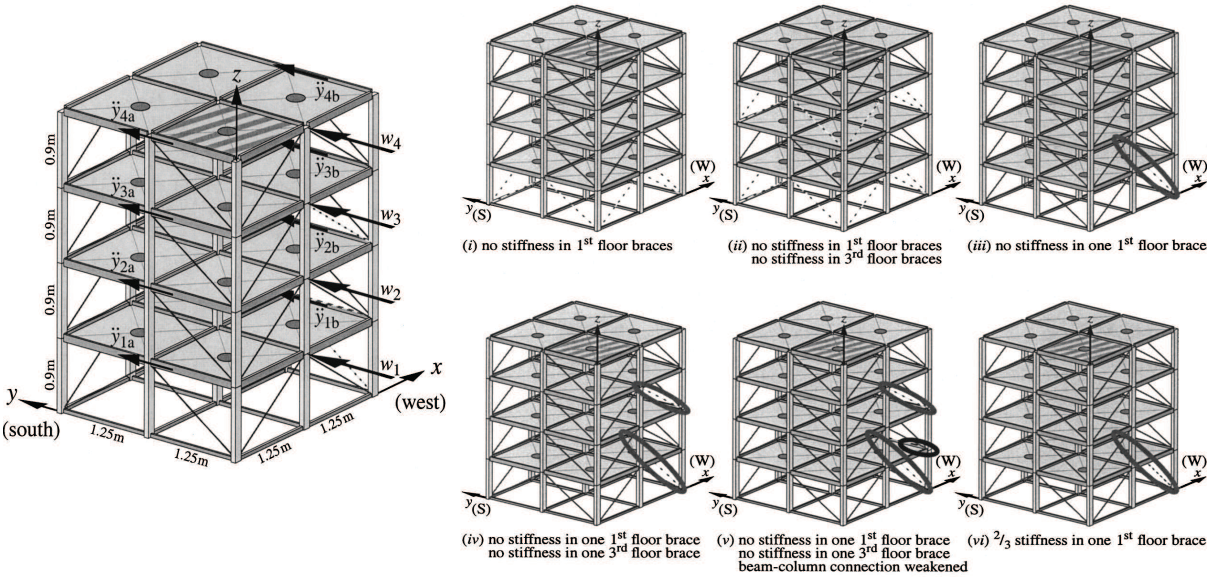
The present work focuses on the development of a vibration-based SHM framework for damage classification in civil and mechanical structures (e.g., the target domain) by cleverly tapping from rich prediction and classification models that were trained on large datasets, such as audio data, largely used in speech and speaker recognition problems (the source domain). The common ground between these two domains is that they are both based on vibrating systems. It is possible to learn from a source domain, not directly related to the structural problem but with an enormously rich database, and transfer this knowledge to the target domain, based on the shared underlying physics (on vibrating systems and their time-frequency features). The transfer process is facilitated by the adoption of Mel Frequency Cepstral Coefficients as features [12, 13]. These coefficients are vastly used in speaker recognition tasks, and their use as DSFs in structural damage assessment should facilitate the transition between the audio and structural domains and enable an explainable artificial intelligence.

DATA

Audio Domain. In the following work, the audio records collection adopted as a source domain is the VoxCeleb dataset. It is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. It consists of records belonging to more than 7000 speakers, where each audio-video segment is at least 3 seconds long for a total of more than 2000 hours of recorded data. Only the audio component of the dataset is considered in this study, while the videos are discarded. The dataset was released in two stages, as VoxCeleb1 and VoxCeleb2 [14]. VoxCeleb1 contains over 100,000 utterances for 1251 celebrities, while VoxCeleb2 contains over 1 million utterances from over 6000 celebrities extracted from videos uploaded to YouTube. The datasets are fairly gender-balanced, (VoxCeleb1 - 55% male, VoxCeleb2 - 61% male). The speakers span a wide range of different ethnicities, accents, professions, and ages. The records are characterized by corruption with real-world noise, consisting of background chatter, laughter, overlapping speech, room acoustics, and a range in the quality of recording equipment and channel noise.

Structural Domain. The target domain is represented by the dataset obtained from numerical simulations of the 12DOF shear-type building model used as benchmark model from the IASC-ASCE SHM Task Group [15]. The response of the system is simulated under seven (one undamaged and six damaged) structural conditions, considering different levels of damage (from heavy damage such as no stiffness in any of the braces of the first and third stories (damage case D2) to light damage, e.g. 33% stiffness loss in one brace (damage case D6). The reader is referred to [15] for a more detailed description of the simulated system and to Figure1 for the description of the damaged scenarios.

The simulation of the system response is implemented considering 16 sensor locations. The target domain dataset comprises 35200 and 9600 simulated time-histories for the undamaged and damaged scenarios respectively (100 simulations for each damage condition). The disparity in the number of damaged and undamaged realizations (undamaged - 78.6%, damaged - 21.4%) mimics the unbalanced nature that commonly characterizes the measurement collection for civil and mechanical monitored systems. Each measurement realization is 10 minutes long and is sampled at 0.0025 s.



(*i*) no stiffness in 1st floor braces
(*ii*) no stiffness in 1st floor braces no stiffness in 3rd floor braces
(*iii*) no stiffness in one 1st floor brace
(*iv*) no stiffness in one 1st floor brace no stiffness in one 3rd floor brace
(*v*) no stiffness in one 1st floor brace no stiffness in one 3rd floor brace beam-column connection weakened
(*vi*) 2/3 stiffness in one 1st floor brace

Figure 1: Diagram of 12DOF analytical model and of the six damage scenarios. The w_i are excitations and the $\ddot{y}_{ij}$ are accelerometer measurements in the y-direction; (the ones in the x-direction are omitted for clarity).

ANALYSIS

The proposed procedure relies on creating a model trained partially on the audio records of the source domain and partly on the structural vibration response of the structure. The model will learn the higher-level features characterizing vibration records from the rich audio dataset and then specialize its knowledge on the chosen structural dataset. The goal is to enrich the model's ability to discriminate between classes on the audio records, presenting multiple different classes with more information to learn. The same model, trained only on the structural dataset, would be ill-conditioned since these datasets tend to be unbalanced in the number of records available for the different classes and would not provide a robust and reliable collection of information.

Two different types of classifications are addressed in this work. First, a binary classification is carried on, where the final aim is to label the tested records from the structure system as either "damaged" or "undamaged". Then a multiclass classification is exploited where the goal is to distinguish between the undamaged condition and the different types of damage scenarios. The proposed TL strategy consisting of four steps is addressed: 1. the Data Preparation; 2. the Feature Extractions; 3. the Training Phase; 4. the Test Phase.

Data Preparation. In the data preparation stage, the source domain and the target domain datasets need to be properly defined and uniformed to avoid ill-conditioned relations between the two domains. In this stage, the train and test datasets are constructed, where the first collects data from the audio dataset and a section from the structural systems, while the test set consists of data only from the structural system.

It is fundamental to highlight how the target domain and the source domain present different frequency content. Generally, audio records show a frequency content that can space from low-frequency value up to kHz, and the sampling frequency usually adopted to collect those records is very high (8 kHz or 16kHz). Lower frequency contents characterize structural systems that rarely exceed the 20-30 Hz range, and consequently, the needed sampling frequency is much smaller (100-200 Hz) with respect to the one used for audio records. This discrepancy in frequency content requires a domain adaptation procedure to uniform the frequency spectrum of the source and target domains. In this work, the structural system presents records sampled at 400 Hz, while the VoxCeleb utterances are recorded at 16 kHz. To achieve an equivalence in terms of frequency content between the domains, a transformation procedure is implemented to bring the structural frequency content to the same frequency range of the audio records (Figure 2).

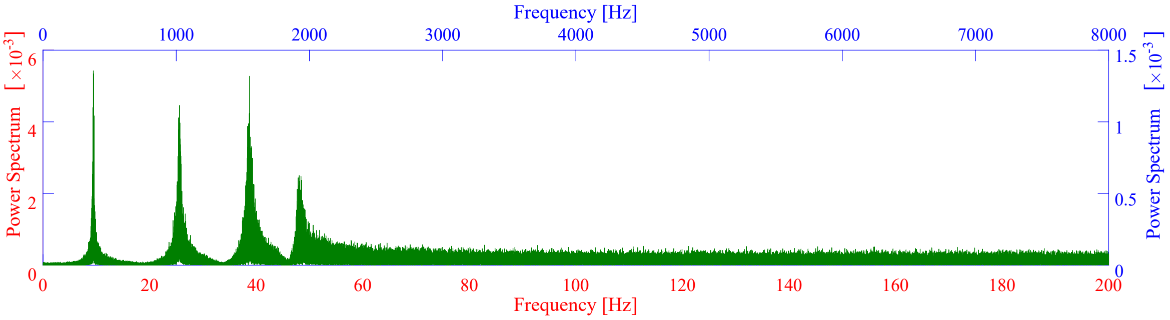


Figure 2: Power Spectrum of an undamaged record for the first sensor. In red the original frequency scale of the structural dataset, in blue the modified scale obtained after frequency domain transformation.

To build a richer model, both the audio and structural datasets have been augmented. In the VoxCeleb dataset, the records are augmented with reverberation, noise, music, and babble using the MUSAN dataset. Once the corrupted records are created, they are combined with the original clean data so that the trained model can learn and discriminate between cleaned signals and signals with noise and disturbances. For the structural dataset, the original raw signal for the numerical simulation is corrupted with white Gaussian noise characterized by 10% RMS, and in each simulation, a small perturbation on the mass and stiffness of each degree of freedom is introduced.

The dataset adopted in the training phase consists of two groups: the first group collects the audio features to train the first part of the model, and the second group consists of records belonging to the structural system. The VoxCeleb dataset will be used entirely for the training, while for the structural systems, only 80% of the records will be adopted for the training. The remaining 20% of the structural data for the creation of the test set.

Feature extraction. In the present work, the Mel Frequency Cepstral Coefficients [12] are adopted as damage sensitive features. For both the VoxCeleb audio domain and for the structural target domain, which is properly transformed, the MFCC vector consists of 30 coefficients per frame. Besides the Mel Frequency Cepstral Coefficients, another set of features is considered and added into the feature vector used to train the classification model, pitch, delta-pitch, and probability of voicing features. Pitch is a perceived quantity related to the fundamental frequency of vibration of the system to which it is referred to. The final damage sensitive feature vector adopted is a 33-dimensional vector. The extraction process is implemented for every record collected in the audio dataset and structural measurements for the training set, and the structural records collected in the test dataset.

Training phase. The features extracted in the previous step are prepared for the training process. They are randomly selected and assigned to the proper tags of classes, creating the dataset, which will be the input for a neural network model. This process is done both for the structural dataset and the audio set. Then, the cepstral mean normalization procedure is applied to the features to make them all zero-mean and remove the convolved noise within the signal. Additionally, for the audio dataset, the possible silence frames are removed. In [16], the authors show that training a PLDA classifier on fixed-length embeddings extracted from the higher layers of a speaker recognition TDNN (which they refer to as "x-vectors") achieves superior performance on out-of-class speaker recognition. [17] successfully implement a similar strategy for automated emotion detection in speech. Following an equivalent approach, the classification task is implemented here, assuming that such a network learns dense representations of speech segments in its upper layers and that these abstract information can be used to classify later the structural health condition of the 12DOF system. The features are derived from the audio domain and used to pre-train a TDNN architecture on a speaker recognition task [12]. Here, the same 9-layer architecture and training methodology adopted in [16] is implemented, using the training script published as part of the Kaldi toolkit [18]. Time-Delay Neural Network is a multilayer artificial neural network architecture able to capture an unknown system's dynamics by modeling a flexible-structured network that will imitate the system by adaptively changing its parameters. This architecture maps a finite time sequence into a single output. Each layer of a TDNN processes a context window from the previous layer, which means that lower layers will have a smaller receptive field and therefore model local features, and higher layers will have a bigger receptive field and thus model long-term dependencies from the slice of features. The input features are 33-dimensional filterbanks with a frame-length of 25ms, mean-normalized over a sliding window of up to 3 seconds.

The TDNN is trained to classify the N classes in the training data. In the audio features, those N classes represent the speakers. In the structural monitoring case, those classes will be two for the binary classification (damaged and undamaged) or N representing different damage classes for the multiclass classification task. In a second stage of the training process, once the presented TDNN architecture is trained using the VoxCeleb audio dataset, it is possible to extract the higher-level features, the x-vectors embedded in the network's intermediate layers. Out of the 9th layer network, the 6th layer is extracted as an x-vector, where the basic information that makes a speaker is learned. These features collect generic information regarding the source domain and characterize a classification process based on the frequency content of signals, which is not strictly dependent on the audio domain. After extracting the x-vector, the mean feature vector is computed, and it is removed for centering the evaluation x-vectors. Starting from the feature x-vector, the trained ability of the model to separate classes is maximized in the final stage in our pipeline with the application of a LDA/PLDA [16, 19] strategy. First, the feature vector is projected into a lower-dimensional G features space adopting the Linear Discriminant Analysis (LDA) approach, which identifies the subspace where the data between different classes is most spread out, relative to the spread within each category. Then, to build an accurate probability model for the new unknown class and find the likelihood estimation, a PLDA model is trained, which takes that projected space derived in the LDA step and rotate it and shift it in a way that stays in the G dimensional space. The PLDA model assumes based on the priors probabilities associated with the LDA analysis and sets an actual probability distribution.

Test phase. In the test phase, the model's input is represented by the features extracted from the test set, and the output is the log-likelihood ratio or similarity score, s, given for each analyzed record and for each class, which says how close each record is to one of the classes. The record is assigned to the class with the higher score. The evaluation data protocol comprises a list of trials, each corresponding to a pair of structural records. The accuracy of the classification procedures is evaluated using the same verification approach adopted for the speaker recognition task, where in the damage detection, it is verified if the tested record matches or not the target health condition. At the end, the performance of the classification model will be given in terms of the Equal Error Rate (EER) or in terms of accuracy. Equal error rate is used to determine the threshold value for a system when its false acceptance rate (FAR) and false rejection rate (FRR) are equal. The accuracy is computed instead as the ratio between the total number of true positive and true negative cases and the total number of inspected cases. The graphic representation of the miss probability by the false alarm rate is Detection Error Trade-Off (DET) Curve [12].

RESULTS

The results are addressed in Table1 in terms of EER and accuracy for the two classification tasks: a binary classification between undamaged and damaged conditions; a multiclass classification to differentiate the classes according to the damage type. The two tasks are carried on considering two different configurations of the trained model on the VoxCeleb dataset. In the first arrangement, the dataset presents no noise augmentation, while in the second, the MUSAN dataset is adopted to augment the audio dataset.

Initially, the damage detection framework is tested on a dataset (Case A) with well separated classes which collects the data from the undamaged and the first two damage conditions (D1 and D2). These two damage conditions represent the most severe damage conditions and contemplate, respectively, no stiffness in the braces of the first story (D1) and no stiffness in any of the braces of the first and third stories (D2). Subsequently, the framework performance is tested considering the dataset with the undamaged and all the damaged conditions, from D1 to D6, included (Case B). The performance of the two classification tasks is visualized in Figure3 and Figure5, given in terms of the DET curve.

Task	**VoxCeleb Unaugmented**		**VoxCeleb Noise Augmented**	
	EER [%]	Accuracy [%]	EER [%]	Accuracy [%]
Binary Classification (Case A)	3.08	97.15	3.08	96.77
Multiclass Classification (Case A)	3.85	96.02	4.23	95.64
Binary Classification (Case B)	23.13	76.90	21.06	78.99
Multiclass Classification (Case B)	18.32	81.44	17.45	82.47

Table 1: Results for the different classification tasks considering the VoxCeleb dataset with and without the noise augmentation with the MUSAN dataset for Case A and Case B

Audio Multiclass Classification: Speaker Recognition Task.

The preliminary results coming for the implementation of the presented strategy seem to be extremely promising. The first step concerns the investigation of the accuracy of the TDNN model trained on the VoxCeleb dataset. It is important to check this model's ability to achieve the speaker classification task using the previously addressed features, parameters, and structure. The test results show the robustness of the model trained on the source domain in the multiclass classification, with an EER of 3.26%. Additionally, the noise augmentation of the dataset with the MUSAN disturbances shows a beneficial effect in the classification task, leading to an EER of 2.64%. The obtained results showed an improvement in adopting the pitch features and the MFCCs with respect to the original reference formulation of the framework, based only on the MFCCs, which lead to an EER of 3.13% considering the augmented dataset.

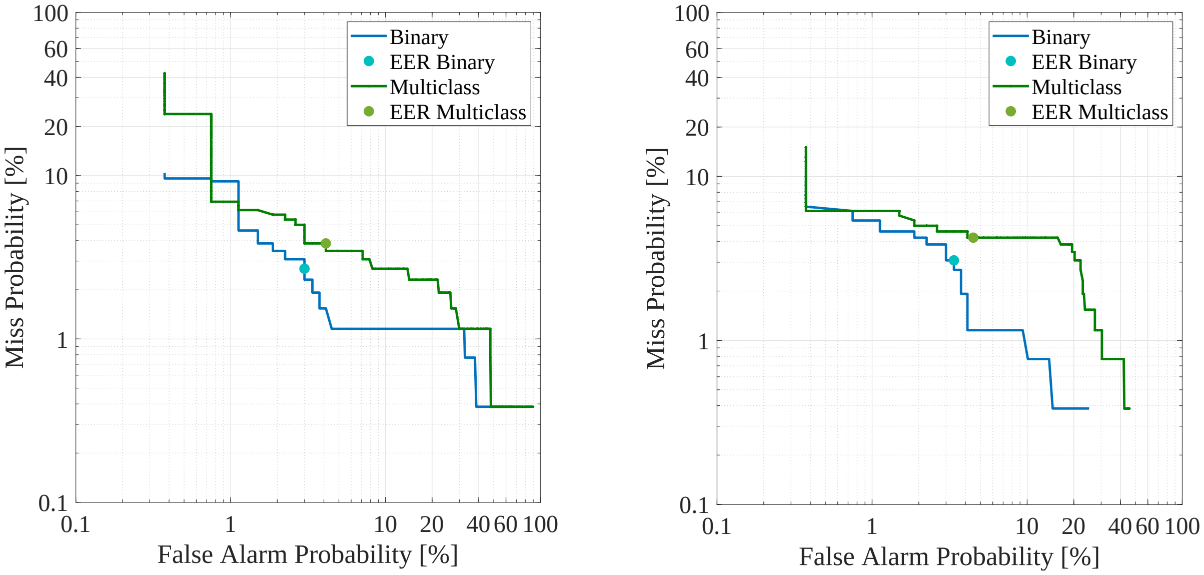
Structural Binary and Multiclass Classification for Case A.

The model trained on the dataset for case A performs with a high accuracy both in the binary and multiclass classification tasks. In this configuration only the undamaged and first two damaged conditions are considered in the training. In the binary classification (Figure 3) the model achieves an EER of 3.08% (LDA dimension=3) considering both the VoxCeleb dataset without noise augmentation and its augmented version. In the multiclass classification (Figure 3) the model achieves an EER of 3.85% and 4.23% (LDA dimension=3) considering, respectively, the VoxCeleb dataset without noise augmentation and its augmented version. It is evident, how in these classification tasks, associated with the dataset in case A, the addition of noise in the audio domain does not lead to a better performance.

Figure 4 presents the LDA-transformed x-vectors for the three LDA dimensions in the augmented VoxCeleb dataset. As expected the damaged and undamaged scenarios are correctly separated in the binary classification, and even when the discrimination gets more granular in the multiclass case, the two damage classes are correctly classified independently. It is evident how the first and second LDA dimensions, which are associated with the biggest eigenvalues, play a key role in separating the x-vectors features related to the two classes.

Structural Binary and Multiclass Classification for Case B.

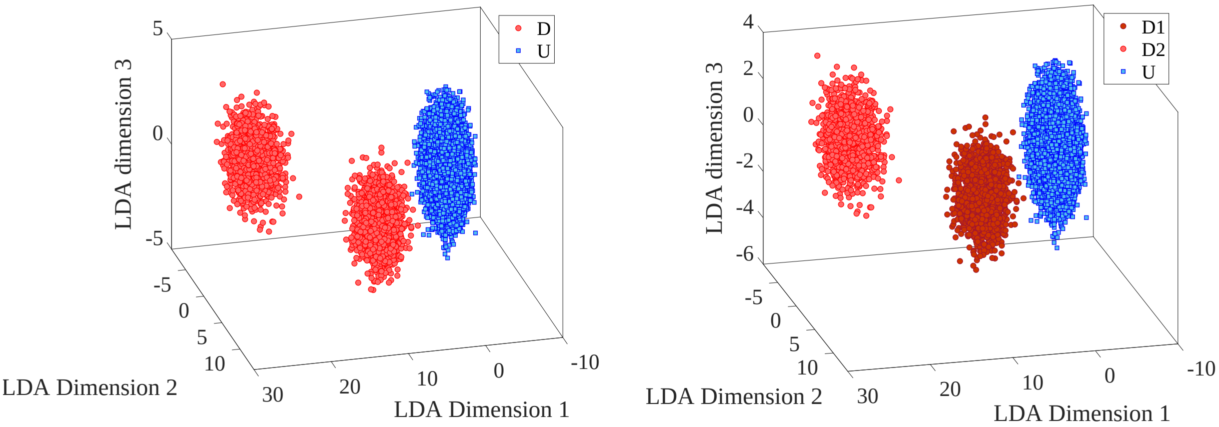
The binary classification task implemented for the dataset in case B, leads to an EER of 23.13% (LDA dimension=5) when considering the VoxCeleb dataset without noise augmentation and to an EER of 21.06% (LDA dimension=6)



(a)

(b)

Figure 3: DET curves for the binary and multiclass tasks, considering the dataset in case A: (a) Unaugmented VoxCeleb dataset; (b) Noise ugmented VoxCeleb dataset.



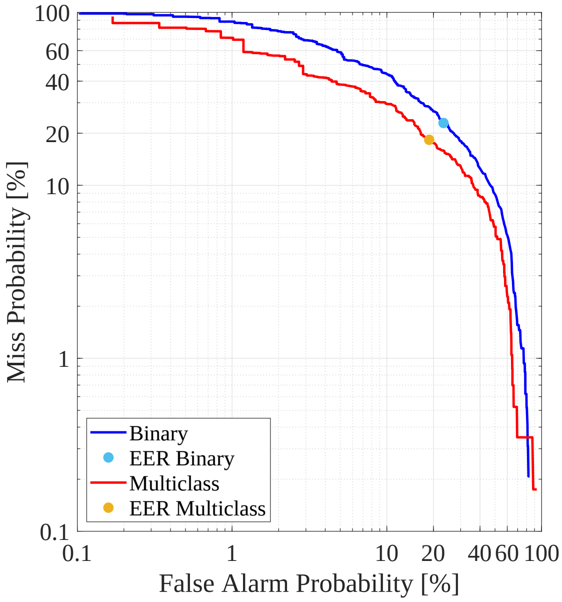
(a)

(b)

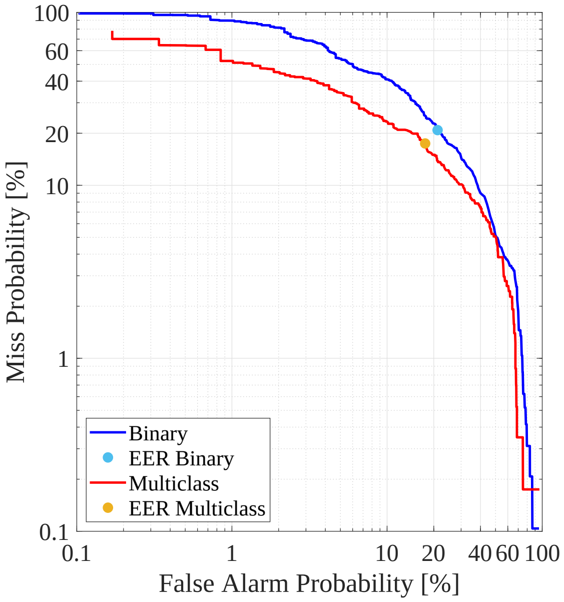
Figure 4: Three components of x-vectors transformed via LDA for the noise augmented VoxCeleb dataset, considering the dataset in case A: (a) binary classification; (b) multiclass classification.

when the MUSAN dataset is considered for augmenting the dataset (Figure 5). The results demonstrate the capability achievable in the classification task by training the TDNN architecture on the audio source. Figure 6 presents the LDA-transformed x-vectors for the first four LDA dimensions in the augmented VoxCeleb dataset. The majority of the cases are correctly classified. However, some of the conditions, belonging to the less severe damage cases (i.e., damage scenarios D6) are wrongly classified. In this case, it is evident how the first LDA dimension, which is associated with the biggest eigenvalue, plays a key role in separating the x-vectors features related to the two classes, while the other features have a low influence.

The multiclass classification task implemented for the dataset in case B, leads to an EER of 18.32% (LDA dimension=5) when considering the VoxCeleb dataset without noise augmentation and to an EER of 17.45% (LDA dimension=3) when the MUSAN dataset is considered for augmenting the dataset (Figure5). Figure 7 shows the LDA-transformed x-vectors for the three LDA dimensions in the case of noise augmented VoxCeleb dataset. Intuitively, the x-vectors features linked with the most severe damages, scenarios D1 and D2, are perfectly separated from the clusters describing the other damaged and undamaged scenarios. The clusters representing the features related to damage D4 and D5, which are less severe failure mechanisms with respect to cases D1 and D2 and

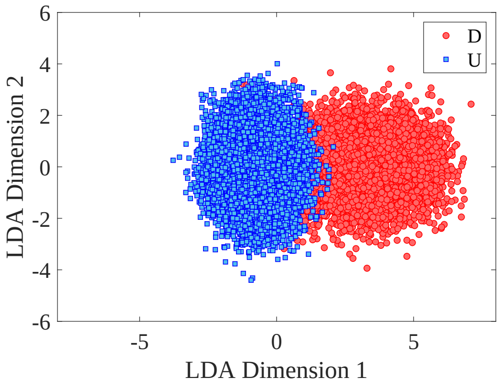


(a)

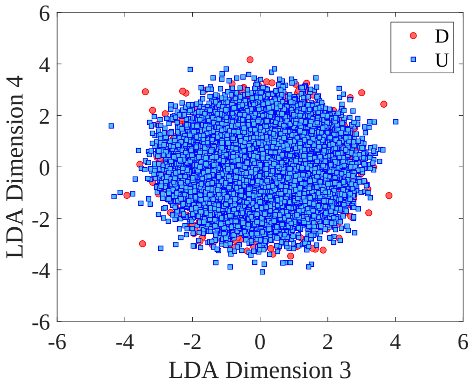


(b)

Figure 5: DET curves for the binary and multiclass tasks, considering the dataset in case B: (a) Unaugmented VoxCeleb dataset(b) Noise ugmented VoxCeleb dataset.



(a)



(b)

Figure 6: First four components of x-vectors transformed via LDA for the binary classification task considering the noise augmented VoxCeleb dataset in case B: (a) LDA Dimension 1 vs. LDA Dimension 2; (b) LDA Dimension 3 vs. LDA Dimension 4.

consist of more localized damage, are still correctly classified. Damage D3, which still interests the loss of stiffness of a whole structural element, is correctly distinguished from the undamaged conditions, even with some wrongly classified cases. On the other hand, damage D6, which corresponds to a partial loss of stiffness for one structural element, is improperly classified.

CONCLUSION

The present study addresses a novel detection SHM strategy based on adopting a transfer learning approach from the audio domain to the structural domain. The framework is based on constructing a richer Probabilistic Linear Discriminant Analysis model starting from the x-vector features extracted from a Time-Delay Neural Network model trained on audio features. The trained model is then used to classify the test data from a structural system in a binary and multiclass classification task.

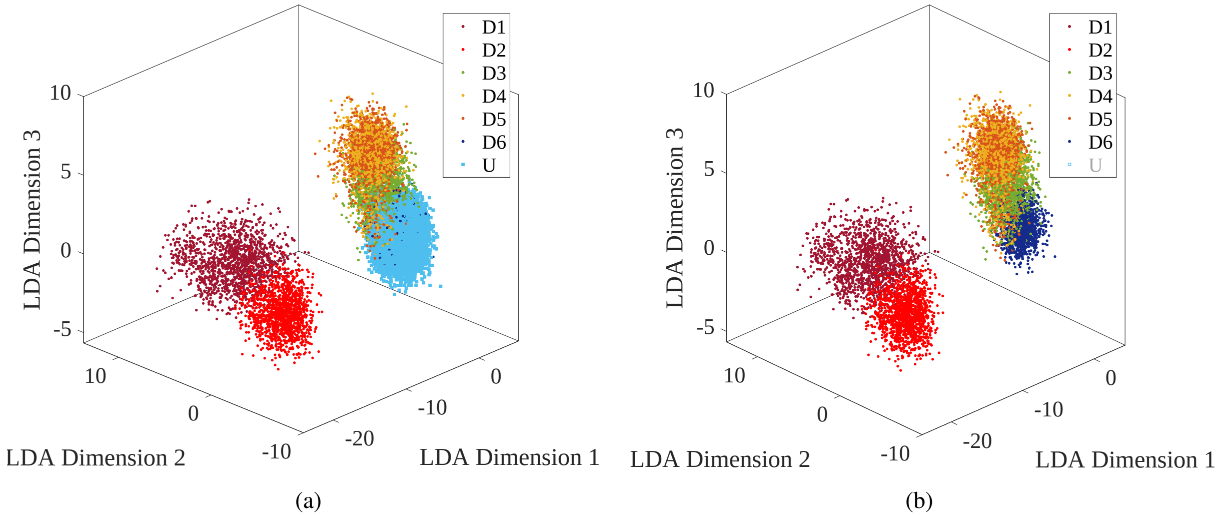


Figure 7: Three components of x-vectors transformed via LDA for the multiclass by damage type classification task considering the noise augmented VoxCeleb dataset in case B: (a) LDA Dimension 1 vs. LDA Dimension 2 for damaged and undamaged scenarios; (b) LDA Dimension 1 vs. LDA Dimension 2 for damaged scenarios.

The proposed methodology is presented and tested to assess damage classification in a simulated 12 DOF shear-type system. The results show strong reliability for the trained model obtained by training the TDNN architecture on the audio source.

The model shows an excellent performance in both binary (damaged and undamaged classification) and multiclass (undamaged and different damaged cases) classification when considering a dataset with very well separated classes (case A). In the analysis related to the second dataset (case B), the EER shows an increase with respect to case A, which is attributable to the miss-classification of the damage classes related to the less severe damages. The outcome for case B demonstrates a highly promising classification ability, showing that, even without any starting information about the structural target system, the model can correctly achieve the classification tasks starting from the simple similarities characterizing these audio signals and their frequency contents. The corruption of the audio dataset with additional external disturbances, like noise and music, did not impact the final accuracy of the model in case A, while in case B, it helps the classification task both in the binary and multiclass cases.

REFERENCES

[1] Farrar, Charles R and Worden, Keith. *Structural health monitoring: a machine learning perspective*. John Wiley & Sons, 2012.

[2] Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[3] Lu, Jie, Behbood, Vahid, Hao, Peng, Zuo, Hua, Xue, Shan, and Zhang, Guangquan. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[4] Bengio, Yoshua. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[5] Bengio, Yoshua, Bastien, Frédéric, Bergeron, Arnaud, Boulanger-Lewandowski, Nicolas, Breuel, Thomas, Chherawala, Youssouf, Cisse, Moustapha, Côté, Myriam, Erhan, Dumitru, Eustache, Jeremy, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172, 2011.

[6] Caruana, Rich. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[7] Hinton, Geoffrey E, Osindero, Simon, and Teh, Yee-Whye. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[8] Erhan, Dumitru, Bengio, Yoshua, Courville, Aaron, Manzagol, Pierre-Antoine, Vincent, Pascal, and Bengio, Samy. Why does unsupervised pre-training help deep discriminant learning? 2009.

[9] Lee, Honglak, Pham, Peter, Largman, Yan, and Ng, Andrew. Unsupervised feature learning for audio classification using convolutional deep belief networks. *Advances in neural information processing systems*, 22:1096–1104, 2009.

[10] Thomas, Samuel, Seltzer, Michael L, Church, Kenneth, and Hermansky, Hynek. Deep neural network features and semi-supervised training for low resource speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6704–6708. IEEE, 2013.

[11] Veselỳ, Karel, Karafiát, Martin, Grézl, František, Janda, Miloš, and Egorova, Ekaterina. The language-independent bottleneck features. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 336–341. IEEE, 2012.

[12] Beigi, Homayoon. In *Fundamentals of Speaker Recognition*. Springer, 2011.

[13] Balsamo, Luciana, Betti, Raimondo, and Beigi, Homayoon. A structural health monitoring strategy using cepstral features. *Journal of Sound and Vibration*, 333(19):4526–4542, 2014.

[14] Chung, Joon Son, Nagrani, Arsha, and Zisserman, Andrew. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[15] Johnson, Erik A, Lam, Heung-Fai, Katafygiotis, Lambros S, and Beck, James L. Phase i iasc-asce structural health monitoring benchmark problem using simulated data. *Journal of engineering mechanics*, 130(1):3–15, 2004.

[16] Snyder, David, Garcia-Romero, Daniel, Sell, Gregory, Povey, Daniel, and Khudanpur, Sanjeev. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[17] Ananthram, Amith, Saravanakumar, Kailash Karthik, Huynh, Jessica, and Beigi, Homayoon. Multi-modal emotion detection with transfer learning. *arXiv preprint arXiv:2011.07065*, 2020.

[18] Povey, Daniel, Ghoshal, Arnab, Boulianne, Gilles, Burget, Lukas, Glembek, Ondrej, Goel, Nagendra, Hannemann, Mirko, Motlicek, Petr, Qian, Yanmin, Schwarz, Petr, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[19] Ioffe, Sergey. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer, 2006.