Application of Speaker Recognition x-Vectors to Structural Health Monitoring

Kyle L. Hom[1], Homayoon Beigi[2], Raimondo Betti[1]

[1] Department of Civil Engineering and Engineering Mechanics
Columbia University, New York, NY 10027

[2] Recognition Technologies, Inc.

ABSTRACT

The domain overlap between speech and structural vibration presents opportunities to leverage advances in speaker recognition for structural health monitoring. Classification of x-vectors, which are the outputs of a pre-final layer from a time-delay neural network (TDNN) acoustic model, has been used to recent success in speaker discrimination. x-vectors present a flexible speaker representation for increased classification robustness, as they contain intermediate modeling parameters rather than outputs for a specific identification task. In investigation of the parallels between speech and structural acoustics, this paper explores the viability of the x-vector speaker recognition system for structural damage detection. A TDNN following the x-vector structure is trained to classify damage scenarios from the Z24 Bridge Benchmark, using cepstral and pitch features from accelerometer measurements. x-vectors are calculated for each measurement, which are used to train a probabilistic linear discriminant analysis (PLDA) model for Z24 damage scenario categorization. This approach yields strong performance in damage detection and classification. As the x-vector system generates representations of structural damage not strictly particular to the Z24 bridge, x-vectors for measurements from other structures can be generated from the developed TDNN. A new PLDA model per structure can then be trained on each structure's x-vectors to identify their respective damage scenarios. We apply this method to the LANL SHM Alamosa Canyon Bridge and UC-Irvine Bridge Column studies, obtaining strong results in damage detection. This method demonstrates the feasibility of speaker recognition techniques for structural health monitoring, and shows significant potential for output-only structural health assessment.

Keywords: Structural Health Monitoring, Time-Delay Neural Networks, x-vectors, Transfer Learning, Z24 Bridge

INTRODUCTION

Structural health monitoring (SHM) and speaker recognition both leverage hidden information to find dynamic behavior and recognize identifying characteristics of an acoustic system. The challenges unique to SHM are often difficult to disentangle; natural degradation over the life of a structure is often unrecorded until failure, and the necessary sparsity of sensor instrumentation provides limited representation of modified dynamics. Though much work in SHM has focused on constitutive models to improve observation of hidden dynamics, there may be much to be gained from investigating developments in speaker recognition. For example, nonlinear dynamics have been captured by data-driven mapping techniques such as neural networks, i-vectors, and Gaussian mixture models, which are topical to the current research efforts in speaker recognition [1]. In this paper, we introduce a recently developed technique in speaker recognition to improve damage detection for SHM.

As both speech and structural vibrations fall under the acoustic domain, we can use a structural measurement dataset to test speaker recognition techniques. One established resource is the Z24 Bridge Benchmark provided by the KU Leuven Structural Mechanics Section for development of damage detection techniques [2, 3]. Efforts in constitutive modeling have yielded some success in identifying damage for the Z24 Bridge Benchmark, though limited by how slight the changes in spectral features are over the progressive introduction of damage [4, 5]. However, with current integration of techniques from machine learning into conventional structural analysis tools [6], and the success of speaker recognition features in damage identification [7, 8], nonlinear structural behaviors may be better captured via data-driven methods.

To explore this research path, we use Kaldi, a robust open-source speech recognition toolkit [9]. The Kaldi team at Johns

Hopkins University has recently developed the x-vector speaker recognition technique, which is leading the field in speaker classification [10, 11, 12]. x-vectors are the intermediate layer outputs from a time-delay neural network (TDNN) trained to classify speakers from speech audio. TDNNs were first used in 1989 to learn temporal relationships between speech sequences for phoneme recognition [13], and were recently rediscovered as effective and computationally efficient alternatives to other sequence-detecting neural networks, such as RNNs or LSTMs. Once trained as an acoustic model to find these temporal relationships within speech, the TDNN's penultimate layers provide outputs, or embeddings, which represent speaker characteristics. By extracting these embeddings for classification, instead of using the final, discrete output classes of the TDNN, the x-vector formulation can robustly tolerate variations of the speaker's behavior.

As speaker classification is an analogous problem to damage classification in structural health monitoring, this paper applies the x-vector formulation to damage classification for the Z24 Bridge Benchmark to assess the technique's potential value in the field. We present three tasks to this x-vector method: global damage scenario classification using x-vectors, local damage severity classification using the damage relationships learned in the damage scenario task, and application of the global damage scenario x-vector system to identify structural modification for unseen LANL SHM datasets.

DATA PREPARATION

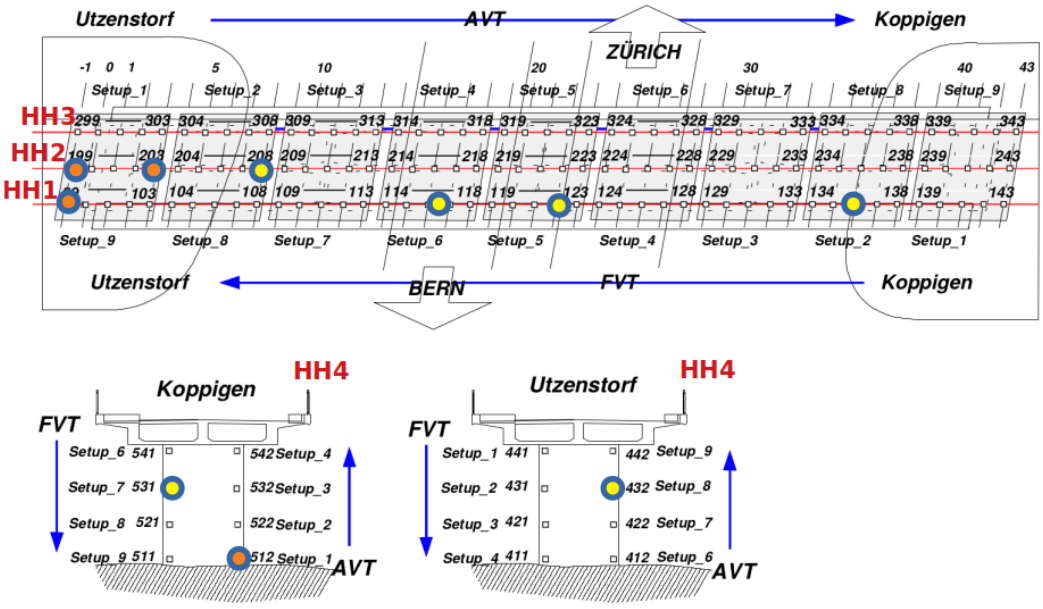


Figure 1: Front and top view of the Z24 bridge instrumentation [14, 3]. Sensor locations reserved from the training dataset R1V, R2(L,V,T), R3V, 208(L,V,T), 432(L,V,T), and 531(L,V,T) are represented by the yellow-filled circles, and sensors with missing data from damage tests 99V, 199L, 203L, 512L, 512V are represented by the orange-filled circles. Both orange and yellow sensors are used in the test set.

Data Resources

We use the Z24 Bridge Benchmark Progressive Damage Test (PDT) provided from the KU Leuven Structural Mechanics Section [2, 3]. The PDT dataset has 17 cases of applied damage (Table 1) for the instrumentation in Figure 1 with ambient and forced vibration testing. The sensor measurements are sampled at 100Hz for 65k samples, for a total of approximately 5,800 minutes of data per damage case. Anomalies in the sensor data are recorded in the test documentation, and are removed or addressed with the recommended processing from the documentation.

The accelerometer measurements must be adjusted so that the frequency content is mapped to the Mel-frequency ranges used for Kaldi's speech features. As the relationship between frequency and Mel-frequency is nearly linear up to 1kHz, we perform a frequency warping of the waveforms to stretch the native sampling rate of 100Hz to 2kHz and the Nyquist frequency from 50Hz to 1kHz. The waveforms are normalized by the largest amplitude in the dataset, to preserve low-amplitude signals from the quantization inherent in conversion to audio file formats. Detrending is performed for all waveforms, as several instances

Table 1: Z24 Progressive Damage Tests and Assigned Local Damage Severity

Damage Case	Estimated Location	Assigned Severity	Damage Scenario	Damage Case	Estimated Location	Assigned Severity	Damage Scenario
1	None	0	Undamaged condition	9	HH1,2,3	3	Spalling of concrete at soffit, 12 m^2
2	HH4	1	Installation of pier settlement system	10	HH1,2,3	3.5	Spalling of concrete at soffit, 24 m^2
3	HH4	2	Lowering of pier, 20 mm	11	Setup1,9	4	Landslide of 1 m at abutment
4	HH4	2	Lowering of pier, 40 mm	12	Setup1,9	4.5	Failure of concrete hinge
5	HH4	3	Lowering of pier, 80 mm	13	Setup1,9	5	Failure of 2 anchor heads
6	HH4	3	Lowering of pier, 95 mm	14	Setup1,9	5.5	Failure of 4 anchor heads
7	None	1	Lifting of pier, tilt of foundation	15	HH1,3	6	Rupture of 2 out of 16 tendons
8	None	1	New reference condition	16	HH1,3	6.5	Rupture of 4 out of 16 tendons
—	—	—	—	17	HH1,3	7	Rupture of 6 out of 16 tendons

of sensor drift are recorded in the Z24 documentation.

Training and Test Data

Assignment of class labels corresponds to the damage scenarios in Table 1. The damage case numbers are assigned to all sensor waveforms from the same damage scenario. Sensor locations for the training set are selected after removing sensor locations which have absent or degraded waveforms during at least one damage scenario, as we want to train the TDNN on a balanced number of waveforms per damage scenario. All removed sensor locations are shown in Figure 1, and we supplement the test set with these sensors' valid waveforms from the other damage scenarios. Hence, the test set in Figure 1 is constructed from the Z24 reference sensor locations (R1, R2, and R3), a 3-DOF sensor opposite R1 (208), two column 3-DOF sensors (432, 531), and the available non-degraded waveforms from the degraded accelerometer set (99V, 199L, 203L, 512L, 512V). These 19 sensor waveforms per scenario in our test set constitute approximately 7% of the 263 total available waveforms per scenario from the Z24 PDT test. Consequently, the training set consists of the remaining 244 sensor waveforms.

Local Damage Labeling

In the interest of leveraging the granularity of the Z24 PDT sensor arrangement, local damage severity ("Estimated Location" and "Assigned Severity") is assigned in Table 1. Severity is labeled as described in the Z24 documentation, from a scale of 1 to 7. We assign severity to sensor measurements at the given Location Label (e.g. only waveforms from HH4 are assigned Severity 1-3 for Cases 2-8). As damage for cases 9-17 are cumulative, severity only increases at the location of induced damage (e.g. waveforms from HH1-3 shifts in Severity from 3.5 to 6 due to Cases 10-15).

ANALYSIS

Data Augmentation

The available sensor data for training a neural network can be increased, or augmented, through speed perturbation and addition of foreground and background noise [12]. Speed variation of 0.9x, 1.0x, and 1.1x is performed before noise corruption. The augmentation is applied with three independent noise sources, consisting of foreground and background white and brown noise, and background traffic recordings from [15]. In total, this augmentation expands the dataset approximately eightfold.

Feature Selection

Mel-Frequency Cepstral Coefficients (MFCCs), pitch, delta-pitch, and probability of voicing features are standard features used in speaker recognition, and are chosen for a preliminary foray into damage detection. Each measurement is divided into evenly-spaced sections, or frames, of 25ms duration sliding over 10ms intervals, over which features are calculated and mean- and variance-normalized.

In Kaldi, the MFCCs are calculated from an inverse discrete cosine transform of the log of the signal's power spectral density

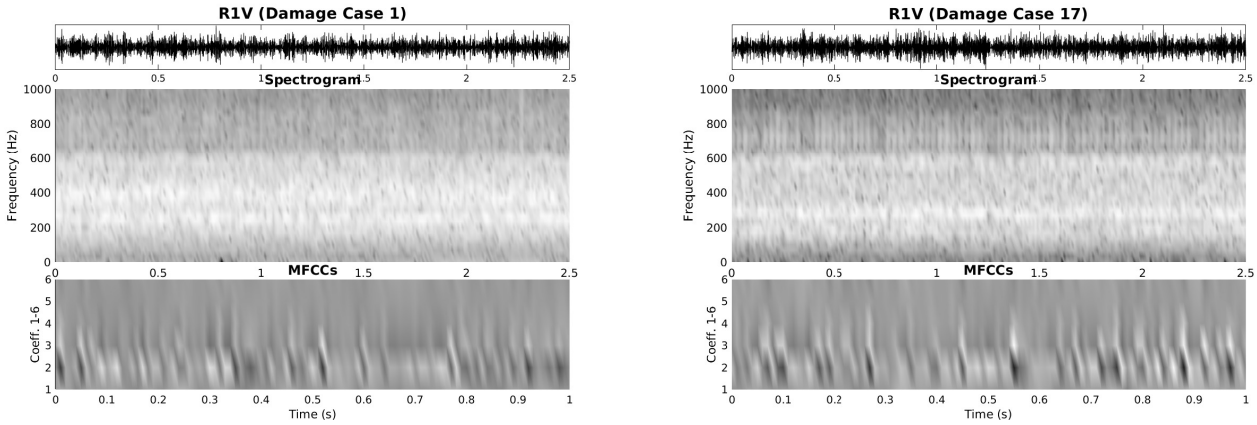
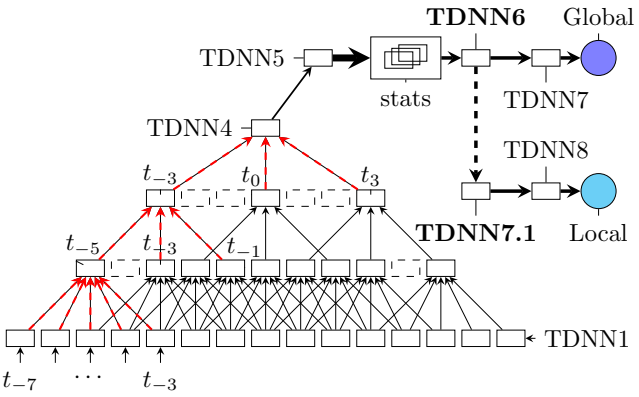


Figure 2: Timeseries, spectrogram, and the first six MFCCs for R1V from damage cases 1 (left) and 17 (right). Though the spectral content has some characteristic banding after the full damage procedure is performed on case 17, the differences between signals is more readily apparent from the MFCC trajectory. In particular, case 17 has sharper and closer bands of MFCCs around durations of 0.0-0.2s and 0.7-1.0s corresponding to the "clicks" and "pops" audible around that periodicity, likely due to the anchor and tendon failures.

[9]. MFCCs represent periodic behavior of frequency spectra over an auditory frequency warping. The dynamics of spectral shifting captured in cepstra is assumed to be of similar utility in the structural domain as it is in the speech domain. As damage is introduced to a structure, modes may shift and new acoustic behaviors may appear, as illustrated in Figure 2. In this study, we assess the effect of varying MFCC resolution on damage detection. We use speech standards of high and low MFCC resolution, or 13 and 30 coefficients respectively. We refer the reader to [1] for a formal description of the calculation of MFCCs.

Normalized pitch, delta-pitch, and probability of voicing features calculated in Kaldi [16] are analogous to a structure's natural frequency, shifts in natural frequency, and the likelihood of such a shift happening. These features are appended to the MFCC features to construct the feature vector per frame, resulting in 16-dimensional and 33-dimensional feature vectors.

Time-Delay Neural Network Architecture



Layer	Layer Context	Total Context	Input × Output
TDNN1	[-2,-1,0,1,2]	5	(N-D×5)×512
TDNN2	[-2,0,2]	9	(512×3)×512
TDNN3	[-3,0,3]	15	(512×3)×512
TDNN4	[0]	15	(512×1)×512
TDNN5	[0]	15	512×1500
stats	[0:T)	T	(1500×T)×3000
TDNN6	[0]	T	3000×512
TDNN7	[0]	T	512×512
Global	[0]	T	512×17
TDNN7.1	[0]	T	512×512
TDNN8	[0]	T	512×512
Local	[0]	T	512×12

Figure 3: 15 frames ($t_{-7} \leq t_0 \leq t_7$) are provided to the input layer (TDNN1). Following the red arrows, TDNN2's t_{-5} node receives 5 frames (corresponding to $t_{-7} \leq t_{-5} \leq t_{-3}$) of N-D features (where $N = 16$ or $N = 33$) as inputs. The TDNN for global damage scenario classification follows the standard x-vector configuration, while the TDNN for local damage severity appends two additional layers (TDNN7.1, TDNN8) to the TDNN6 layer and is retrained on the local severity labels. Bold indicates layers for intermediate output extraction.

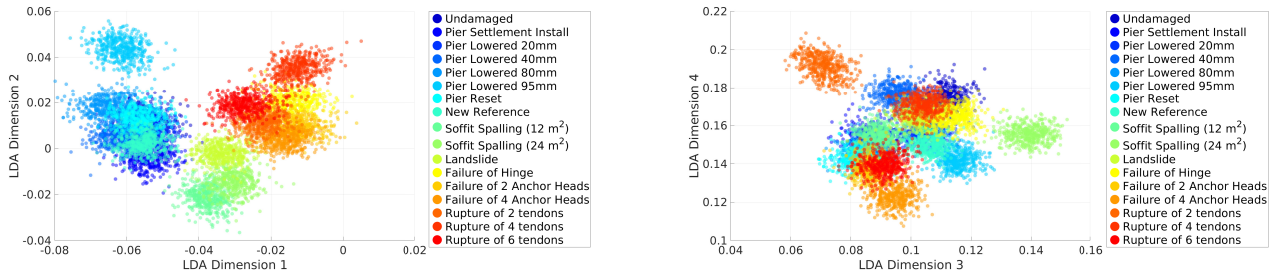
We use a TDNN with the same structure presented in the x-vector formulation [12] to find distinguishing structural damage characteristics across the Z24 PDT scenarios. The TDNN captures vibration dynamics by learning relationships between

features over a sequence of frames corresponding to the damage scenario. Exposure to past or future features from a given point in the sequence is accomplished by defining a range of frames, or context, that are connected in a layer. As the input layer is restricted to an ordered sequence of time-dependent features, the outputs of the following hidden layer is a compressed representation of structural damage dynamics over the defined context.

As shown in Figure 3, the first four layers (TDNN1-TDNN4) collect several frames of context from the sensor waveform before and after the frame being assessed. To reduce sensitivity to the selection of frames when segmenting the signal, a statistics pooling layer is inserted after sufficient frame-level representations are collected. This statistics pooling layer collects all of the TDNN5 outputs in the segment of the measurement input and returns the mean and standard deviation for the segment, compressing the context of the following layers into segment-level representations. The following layers (TDNN6, TDNN7) fit the statistics from the T sequences to the corresponding damage scenarios, and a softmax output layer provides the predicted class of the 17 damage scenarios. This TDNN is trained over eight epochs, and thus sees all training data eight times.

The x-vectors for each accelerometer signal are obtained from the outputs of TDNN6, yielding 512×1 dimensional vectors. We can improve the damage classification performance by attempting to use an intermediate layer's output as the identifier of damage [11, 12]. The final output of the network is supposed to be the best-estimate of the classes, but this is often not the case; variations in the test dataset may be uncaptured by the network, and the output at the end of the network may be inflexible in distinguishing between classes (as the final layer is a softmax output layer, and tends to learn the 'strongest' separations between discrete classes). We hope that extracting outputs at a layer close, but not at the final output layer may yield better class representations and reduce the effect of overfitting the network to particular damage scenarios.

PLDA Classification



(a) Separation of pier lowering, spalling/landslide, and failure with first and second LDA dimensions

(b) Delineation of pier lowering depth with third LDA dimension, failure severity with fourth LDA dimension

Figure 4: Visualization of the LDA-transformed x-vectors from the first four largest eigenvalues (or LDA dimensions), demonstrating separability of the x-vectors. Separation appears to follow intuitive groupings of the damage scenarios and order of importance for discrimination (where first and second eigenvalues correspond to damage mechanisms, and third and fourth to levels of severity). The PLDA technique finds the centers and covariances of these clusters, and provides a log-likelihood ratio for class membership scoring of a presented test example.

We apply a LDA transformation to the x-vectors for dimensionality reduction from 512 to 200 dimensions. To perform classification in the LDA-transformed space, a Probabilistic LDA (PLDA) classifier [17] is developed from the training set of x-vectors and the LDA transform. The PLDA technique assigns continuous probabilities to the classes used in LDA, and so can be used to determine log-likelihood ratios for the test set's membership in the probability distributions of the damage scenarios, within the LDA-transformed space. Hence, where LDA provides the 'primary directions' to maximize between-class scatter and minimize within-class scatter, PLDA provides the probability an example participates in each of these 'primary directions'. Classification is then performed based on the log-likelihood ratios, scoring a test example against the clusters of training examples (as visualized in Figure 4).

Local Damage Identification

Given the spatial granularity of the Z24 PDT sensor grid, we attempt to identify local damage severity over the damage scenarios, utilizing the learned structural acoustic relationships from the previous TDNN. Following the transfer learning technique

from [18, 19, 20], we initialize a new network with the pretrained Z24 damage scenario TDNN's TDNN1-TDNN6 layers, and append new layers TDNN7.1, TDNN8 and a softmax output layer as shown in Figure 3. This local damage TDNN is trained on the 12 local damage severity labels from Table 1 for only three epochs, as less training is required after initializing with the pretrained global damage TDNN. We then follow the same LDA/PLDA classification procedure, extracting local damage embeddings at TDNN7.1.

Validation on LANL Datasets

Though the Z24 damage scenario TDNN's output layer is fitted to the Z24 PDT's 17 damage scenarios, we can use this TDNN to provide x-vectors for other structures. Using the Los Alamos National Lab SHM Experimental Datasets for the Alamosa Canyon Bridge and UCI Column tests [21, 22], we perform the aforementioned data preparation procedure and feed the converted measurements forward through the TDNN. The resulting x-vectors are then classified via the PLDA procedure to assess how well they distinguish between structural modifications. For the Alamosa Bridge, we attempt to detect the addition of a stiffener placed on the midspan of the bridge. For the UCI Column tests, we attempt to distinguish between six cases of progressive loading for two columns with different reinforcement techniques.

RESULTS

Table 2: Damage Classification via x-Vector Approach

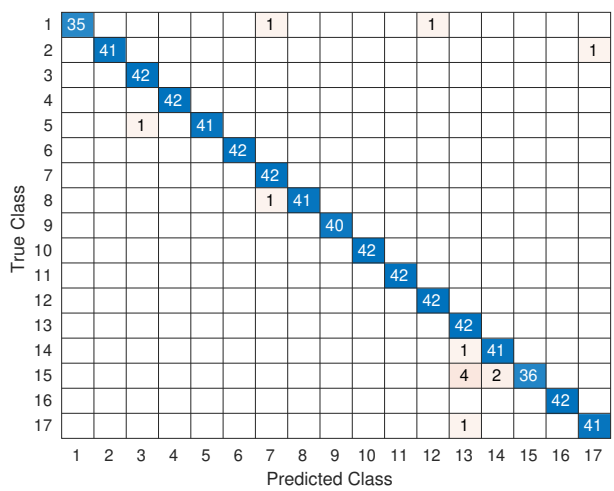
			13 MFCCs, 3 Pitch			30 MFCCs, 3 Pitch		
		Augmentation	Acc.(%)	EER(%)	AUC	Acc.(%)	EER(%)	AUC
Z24 Global	x-vector	—	87.977	5.938	0.989	**91.372**	**3.919**	**0.994**
	x-vector	(aug.)	83.451	7.525	0.983	88.967	5.658	0.986
	softmax	—	97.171	0.643	0.999	**98.161**	**0.455**	**0.999**
	softmax	(aug.)	95.757	1.225	0.999	96.322	1.067	0.999
Z24 Local	TDNN7.1	—	71.146	11.376	0.954	**72.419**	**12.414**	**0.951**
	TDNN7.1	(aug.)	63.366	14.120	0.929	66.195	13.626	0.932
	softmax	—	80.057	6.909	0.969	82.885	6.376	0.9709
	softmax	(aug.)	85.572	6.244	0.977	**85.997**	**5.741**	**0.979**
Alamosa	x-vector	—	**99.972**	**0.023**	**0.999**	99.954	0.046	0.999
	x-vector	(aug.)	99.954	0.037	0.999	99.935	0.065	0.999
UCI Column	x-vector	—	96.005	5.261	0.991	98.707	3.063	0.996
	x-vector	(aug.)	99.745	1.592	0.999	**99.530**	**1.364**	**0.999**

The results for applying the x-vector formulation to the Z24, Alamosa, and UCI datasets are provided in Table 2. We use accuracy, equal-error rate (EER), and area-under-the-curve (AUC) to assess the performance of this technique on these datasets. Our best-performing x-vector damage scenario setup has an EER of 3.919%, which is comparable to the x-vector performance on the VoxCeleb corpus and Speakers in the Wild Core with a minimum EER of 4.16% [12].

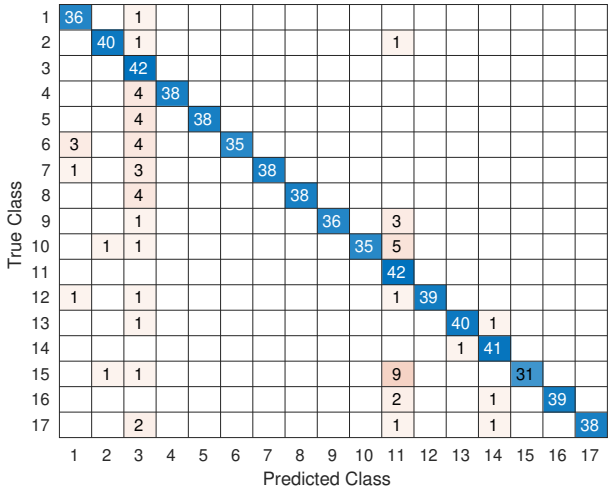
For the global damage scenarios in Figure 5, the softmax output performs best, as it was directly trained to predict the classes. Though the x-vectors do not perform as well in identifying damage type for low-damage scenarios, we see in later results that their flexibility proves beneficial for classifying unseen data.

When classifying local damage severities, the TDNN7.1 output confuses lower severities for higher severities more often than the softmax output in Figure 6. We believe this was caused by the population bias towards lower damage across the sensor locations. This is in part due to the localization of damage methods: for example, tendon failures only increase severity for the sensor locations distributed along the tendon. However, the flexibility of the intermediate layer output classification appears to manifest for prediction of higher severity damage, as the TDNN7.1 output predicts this severity more accurately.

The detection-error-threshold (DET) curves for the damage classification tradeoff space indicate how much we can tolerate

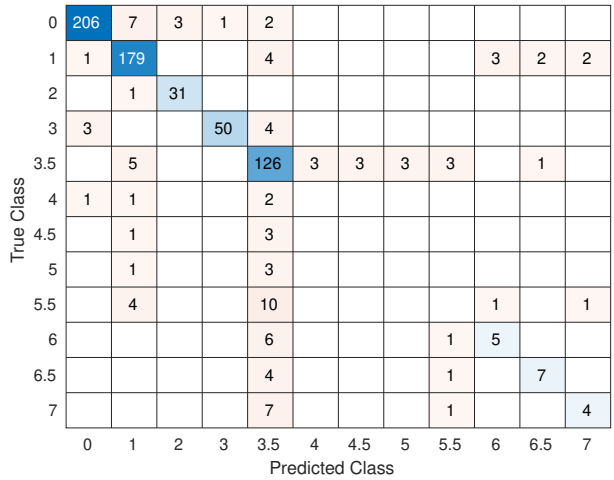


(a) Unaugmented 30 MFCCs, 3 Pitch Softmax Output

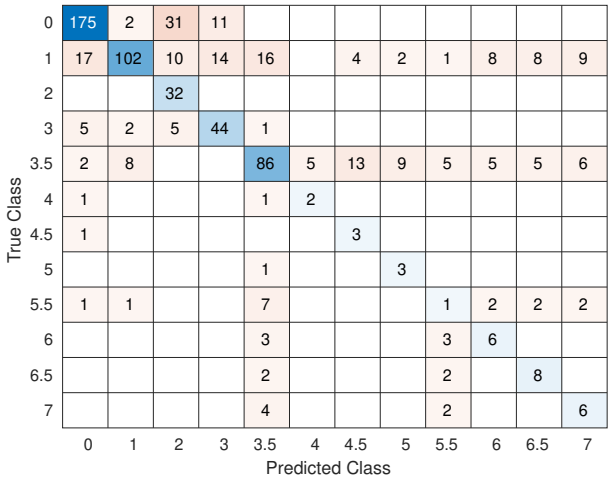


(b) Unaugmented 30 MFCCs, 3 Pitch x-Vector Output

Figure 5: Selected Damage Scenario Confusion Matrices. The x-vector approach appears to be biased towards predicting damage case 3, which is the first instance of pier lowering. This may be a level of damage that is not well associated with increased pier lowering, and may emulate effects of other damage cases.



(a) Augmented 30 MFCCs, 3 Pitch, Softmax Output

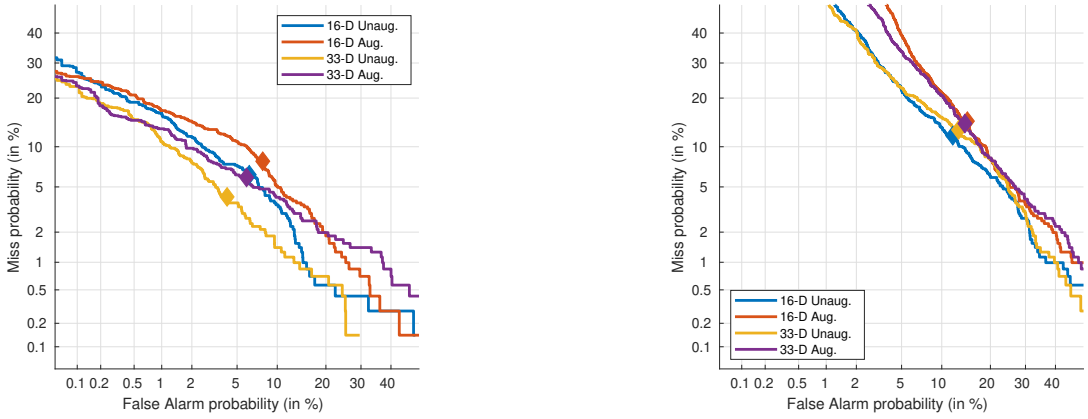


(b) Augmented 30 MFCCs, 3 Pitch, TDNN-7.1 Output

Figure 6: Selected Local Damage Severity Confusion Matrices. The TDNN7.1 output classification performs worse on the lower damage severities, but errs on the side of conservatism in predicting more severe damage at more sensor locations. It correctly identifies severe damage more often than the softmax outputs.

false alarms in exchange for lower detection of damage scenarios [1]. From Figure 7, we observe that we can reduce our miss probability of damage to 1% if we accept a 15% false alarm probability, indicating that the x-vector approach for global damage classification may be effective for use in bridge inspections and damage assessment.

Strong performance is observed from the x-vector classification for the unseen LANL datasets. Of note is the improvement which data augmentation during training of the Z24 TDNN provides for classifying the UCI Column damage: both low-resolution MFCC and high-resolution MFCC feature configurations perform well with data augmentation. Speaker recognition with x-vectors has shown improved performance with training on noise-corrupted and speed-perturbed signals, and it is postulated that this sort of variation in data 'loosens' the fit of the x-vectors to the particular training data [12]. In applying the



(a) Global Damage Scenario

(b) Local Damage Severity

Figure 7: DET Curves for x-Vectors

learned acoustic knowledge from the Z24 dataset in the developed TDNN, this sort of data augmentation may have increased the flexibility of the network in separating damage types across different structures.

CONCLUSIONS

The application of x-vectors for structural damage recognition has been investigated using the Z24 Bridge Benchmark and the LANL SHM Dataset. Damage scenario and local damage identification were performed, through development of two TDNNs for these tasks and PLDA classification of their intermediate outputs. Several feature enhancement techniques were explored to assess impact on damage assessment performance of these methods, including MFCC resolution and noise augmentation. These acoustic representations replicated the success of their parent models for speaker recognition in application to structural damage detection, with very strong results in distinguishing between damage scenarios across various SHM datasets. x-vectors drawn through the developed Z24 damage scenario TDNN demonstrate high potential to be flexible and accurate damage detection features in diagnosing structural health.

In particular, the extension of damage scenario classification to local damage at particular sensors presents a promising field for investigation. As this approach does not take full advantage of the physical interactions between adjacent sensors, potential application of constitutive modeling may contribute physics-based structure to the demonstrated adaptation to empirical data of the x-vector framework. Future work in this area will encompass merging data-driven approaches (such as those presented in this paper) with constitutive analysis methods, and extension of this technique to assess bridges with sparse instrumentation.

ACKNOWLEDGMENTS

The authors would like to thank the KU Leuven Structural Mechanics Section for providing the Z24 Bridge Benchmark dataset, and the Los Alamos National Laboratory for providing the Alamosa Canyon Bridge and UCI Bridge Column datasets.

REFERENCES

[1] Beigi, Homayoon. *Fundamentals of Speaker Recognition*. New York: Springer, Dec. 2011. ISBN: 978-0-387-77591-3. DOI: 10.1007/978-0-387-77592-0.

[2] Maeck, J. and De Roeck, Guido. “Description of Z24 benchmark”. *Mechanical Systems and Signal Processing - MECH SYST SIGNAL PROCESS* 17 (Jan. 2003), pp. 127–131. DOI: 10.1006/mssp.2002.1548.

[3] Reynders, Edwin and De Roeck, Guido. "Continuous Vibration Monitoring and Progressive Damage Testing on the Z24 Bridge". *Encyclopedia of Structural Health Monitoring*. John Wiley and Sons, 2009. ISBN: 9780470061626. DOI: 10.1002/9780470061626.shm165.

[4] Masciotta, Maria-Giovanna et al. "A spectrum-driven damage identification technique: Application and validation through the numerical simulation of the Z24 Bridge". *Mechanical Systems and Signal Processing* 70-71 (2016), pp. 578–600. ISSN: 0888-3270. DOI: https://doi.org/10.1016/j.ymssp.2015.08.027.

[5] Masciotta, Maria, Ramos, Luís, Lourenco, Paulo, and Vasta, Marcello. "Damage Detection on the Z24 Bridge by a Spectral-Based Dynamic Identification Technique". *Dynamics of Civil Structures, Volume 4*. Cham: Springer International Publishing, 2014, pp. 197–206. DOI: 10.1007/978-3-319-04546-7_23.

[6] Azimi, Mohsen, Eslamlou, Armin Dadras, and Pekcan, Gokhan. "Data-Driven Structural Health Monitoring and Damage Detection through Deep Learning: State-of-the-Art Review". *Sensors* 20.10 (2020). ISSN: 1424-8220. DOI: 10.3390/s20102778.

[7] Balsamo, Luciana, Betti, Raimondo, and Beigi, Homayoon. "A structural health monitoring strategy using cepstral features". *Journal of Sound and Vibration* 333 (Sept. 2014), pp. 4526–4542. DOI: 10.1016/j.jsv.2014.04.062.

[8] Balsamo, Luciana, Betti, Raimondo, and Beigi, Homayoon. "Damage Detection Using Large-Scale Covariance Matrix". Vol. 5. Feb. 2014. DOI: 10.1007/978-3-319-04570-2__10.

[9] Povey, Daniel et al. "The Kaldi Speech Recognition Toolkit". *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[10] Snyder, David, Garcia-Romero, Daniel, Povey, Daniel, and Khudanpur, Sanjeev. "Deep Neural Network Embeddings for Text-Independent Speaker Verification". *Proc. Interspeech 2017*. 2017, pp. 999–1003. DOI: 10.21437/Interspeech.2017-620.

[11] Snyder, David et al. "Spoken Language Recognition using X-vectors". June 2018, pp. 105–111. DOI: 10.21437/Odyssey.2018-15.

[12] Snyder, David et al. "X-Vectors: Robust DNN Embeddings for Speaker Recognition". Apr. 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[13] Waibel, Alexander H. et al. "Phoneme recognition using time-delay neural networks". *IEEE Trans. Acoust. Speech Signal Process.* 37 (1989), pp. 328–339.

[14] De Roeck, Guido, Peeters, Bart, and Maeck, Johan. "Dynamic monitoring of civil engineering structures". eng. 2000.

[15] *Soundjay Traffic and Transportation Ambience*. https://www.soundjay.com/transportation-ambience-1.html. Accessed: 2020-05-13.

[16] Ghahremani, Pegah et al. "A pitch extraction algorithm tuned for automatic speech recognition". *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2014, pp. 2494–2498.

[17] Ioffe, Sergey. "Probabilistic Linear Discriminant Analysis". *Computer Vision – ECCV 2006*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542. ISBN: 978-3-540-33839-0.

[18] Ghahremani, P. et al. "Investigation of transfer learning for ASR using LF-MMI trained neural networks". *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2017, pp. 279–286. DOI: 10.1109/ASRU.2017.8268947.

[19] Ananthram, Amith, Saravanakumar, Kailash, Huynh, Jessica, and Beigi, Homayoon. "Multi-Modal Emotion Detection with Transfer Learning". *Natural Language, Dialog and Speech Symposium (NDS2020)*. New York Academy of Science, New York City, USA, Nov. 2020. DOI: 10.13140/RG.2.2.31373.97760.

[20] Zhou, Sitong and Beigi, Homayoon. *A Transfer Learning Method for Speech Emotion Recognition from Automatic Speech Recognition*. Tech. rep. Recognition Technologies Technical Report No. RTI-2020330-01. Mar. 2020.

[21] Farrar, Charles, Cornwell, Phillip, Doebling, Scott, and Prime, Michael. "Structural Health Monitoring Studies of the Alamosa Canyon and I-40 Bridges" (July 2000). DOI: 10.2172/766805.

[22] Farrar, Charles et al. "Damage Identification With Linear Discriminant Operators" (July 2003).