Combining acoustic features and medical data in deep learning networks for voice pathology classification

Ioanna Miliaresi
Department of Informatics
University of Piraeus
Piraeus, Greece
imiliaresi@unipi.gr

Kyriakos Poutos
Department of Informatics
University of Piraeus
Piraeus, Greece
kyriakos@unipi.gr

Aggelos Pikrakis
Department of Informatics
University of Piraeus
Piraeus, Greece
pikrakis@unipi.gr

***Abstract*—In this paper, we present a study on the efficiency of neural networks for the hard problem of automatically classifying voice disorders. To this end, convolutional architectures combined with feed-forward neural networks are used for the classification of four types of voice disorders. Speech signals and data from medical records, collected by the Far Eastern Memorial Hospital (FEMH), involving four speech pathologies, (functional dysphonia, phonotrauma, laryngeal neoplasm and unilateral vocal paralysis), were analyzed and the proposed method participated at the FEMH Voice Data challenge 2019. The respective classification accuracy at the challenge's testing dataset was 57% and the method ranked fifth with a small performance margin from the leading method.**

***Index Terms*—FEMH, Voice disorders, Neoplasm, Phonotrauma, Functional Dysphonia, Vocal Palsy, Neural networks**

I. Introduction

Speech pathology classification refers to the problem of automatically deciding which type of pathology is present in a given recording out of a set of predefined classes of pathology. In the literature, this problem has been traditionally approached from the perspectives of pattern recognition and statistical learning. Nowadays, as deep learning provides state of the art methods in many application fields, a natural need emerges to conduct a deep learning study in the field of automatic voice pathology classification and this is also the goal of this paper.

It is worth noting that in pathological voice classification, the most widely used dataset has so far been the Massachusetts Eye&Ear Infirmary (MEEI) Voice Disorders Database, commercialized by KayPENTAX Corp., where excellent results have been reported with conventional classifiers. As it is pointed in [1], the normal and dysphonic sustained vowels of the KayPENTAX database are actually perfectly separable, but working in the direction of generalization, tests on new datasets have to be performed, where such high detection and classification results may not be possible.

In the context of the 2019 "IEEE BigData Cup Challenges", the FEMH voice data Challenge addressed a 4-class classification problem for recordings stemming from four categories of disorder, namely functional dysphonia, phonotrauma, laryngeal neoplasm and vocal paralysis. The participants were provided with a training databases and the ranking of submitted methods was computed on an undisclosed testing set. The competition was an extension of the 2018 challenge, where the objective was to distinguish healthy from pathological cases and perform classification based on three voice pathologies, namely neoplasm, phonotrauma and vocal palsy. Therefore, the 2019 challenge addressed a harder task, for which there do not exist previous reports in the bibliography. The difficulty of the task at hand is verified by the low classification accuracy of the top five methods. Tables I and II summarize the related work in the field, as it is reflected in the 2018 FEMH challenge. Even though the 2018 challenge addressed a somewhat simpler task, the related methods clearly indicate a preference for mainstream machine learning approaches. It is worth noting that algorithmic details related to the methods participating in the 2019 challenge were not publicly available at the time of writing this paper.

Based on the announced results, our deep learning method ranked fifth, with less than 7 percentage units separating it from the best performing approach. We use deep learning design principles to fuse different information modalities into a single architecture which is trained with the backpropagation principle. Method [2] in Table II also uses deep neural networks to combine acoustic and medical data but in a very different setting. Specifically, two standard feed-forward neural networks are trained and their softmax decisions are combined in a third network. The feature extraction stage includes a Gaussian Mixture Model (GMM) to represent an audio recording statistically before feeding the GMM output to one of the two neural networks. Our approach is very different because we operate on mid-term segments, create 2-D representations, use a modern convolutional architecture for one of the two sub-networks and fuse intermediate learned representations (not softmax vectors) at a last processing layer, i.e., softmax decisions are taken once in the end of the processing chain. This allows for better flow of the error signal during training. Furthermore, we use perturbation features

alongside the medical metadata on the same sub-network.

The next section presents the proposed method, including database description, feature extraction and model architecture. Section III describes the experimental setup, obtained results and discusses some crucial issues. Finally, the last section provides the conclusions obtained from our study.

II. METHOD DESCRIPTION

A. DATASET

The 2019 FEMH dataset consists of speech signals and medical records of patients with four disorders, namely dysphonia, phonotrauma, vocal palsy, functional dysphonia and neoplasm. The speech samples and the related medical records were obtained from a voice clinic in the Far Eastern Memorial Hospital (FEMH). Each medical record contains 34 demographic questions, both categorical and binary, including age, gender, job, habits and symptoms, when the voice became worst, how it happened, whether experienced internal surgery took place or not, how strict the gastroesophageal reflux is and so on. The training dataset contains fifty voice recordings of each disease, where a sustained vowel /a/ is pronounced by pathological speakers. The duration of the recordings lies in the range of 2 to 39 seconds and the sampling frequency varies among recordings. Following the challenge rules, our system's classification efficiency was tested on a testing set of a total of two hundred recordings covering the four classes under study. The testing set details are undisclosed to the time of writing this paper.

Pathological speech usually refers to the condition of speech distortion resulting from atypical ties in voice and/or in the articulatory mechanisms owing to disease, illness or other physical or biological insult to the production system. *Dysphonia* is a disorder of the voice production mechanisms in the larynx with specific perceptual, acoustic, and physical correlates. Dysphonia is actually only one pathology in several different disorders and symptoms, sometimes manifesting as secondary symptom, either as principal. Dysphonia can be organic or functional. Organic dysphonia is due to an anatomical change in the vocal fold, like nodules or benign tumors. Organic dysphonia is due to an anatomical change in the vocal fold, like nodules or benign tumors. Functional dysphonia is assumed when no anatomic changes are known. In this work a variation of functional dysphonia, the hyperfunctional dysphonia is analyzed. Hyperfunctional dysphonia is an excessive involuntary muscle contraction, as a consequence of improper phonation, that results in a hoarse or strained voice. [14].

Vocal hyperfunction refers to (chronic) conditions of abuse and/or misuse of the vocal mechanism due to excessive and/or unbalanced (uncoordinated) muscular forces and is associated with the most frequently occurring types of voice disorders. Vocal hyperfunction is believed to play a primary role in causing the chronic tissue trauma, referred to as *phonotrauma*, that leads to the formation of common vocal-fold lesions (e.g. vocal fold nodules).

Vocal fold paresis/paralysis (palsy) of voice box muscles results in voice changes like hoarseness, breathy voice, extra effort on speaking, excessive air pressure required to produce usual conversational voice and diplophonia (voice sounds like a gargle).

Finally, the term *neoplasm*, refers to various types of cancer, including laryngeal, voice box or vocal cords tumors. Common symptoms are hoarseness, painful swallowing and fatigue.

B. FEATURE EXTRACTION

At the first step of the processing pipeline, a short-term feature extraction method is applied so as to extract a sequence of feature vectors from the time-domain representation of the input signals. The signal is first normalised and re-sampled to 44100Hz. As the input to the first convolutional layer of our architecture requires a fixed-size image and since the recording duration varies, a standard segmentation procedure is adopted in order to generate a sequence of 1.28 s long fixed-length segments per recording. Each such segment is parsed with a 40 ms long moving window with a hop size of 20 ms. At each frame, the Discrete Fourier Transform (DFT) is computed and it is given as input to a mel-filter-bank, with each mel-filter performing a weighted sum of the magnitude of the DFT coefficients that lie inside its frequency range. The logarithm of each filterbank output is then calculated and the discrete cosine transform of the logarithms is computed. As it is standard practice, the first 13 MFCCs are kept and the rest are discarded. Furthermore, to capture the dynamics of the signal, the first-order derivative of the MFCC vector is computed over time and it is appended to the vector of the MFCCs. Finally, the feature vector is augmented with the logarithm of the 26 mel-filterbank outputs, thus resulting into a total of 52 feature values per frame and a sequence of 64 feature vectors per segment. This is treated as a $2-D$ representation (image) of the input segment and it is fed to a convolutional network. In addition, in order to deal with specific impacts of voice pathology to sound quality, the fundamental frequency, jitter and Harmonic to Noise Ratio are computed, using the autocorrelation-based algorithm for periodicity detection in [15]. According to that algorithm, the best candidate for pitch period estimation can be detected by inspecting the position of the maximum of the autocorrelation function of the sound, while the periodicity strength (harmonics-to-noise ratio) is computed from the relative height of this maximum when compared with other peaks.

As it will be shown at the results at Table II, the utilization of the available medical records data is recommended in order to increase the systems classification capability. For every recording, the 34 respective medical measurements are combined with the three aforementioned mid-term segment features (F0, Jitter and HNR) to form a 37×1-dimensional vector that is fed as an input to a separate feed forward network.

Therefore, a novel aspect of our approach is that we are using a network architecture consisting of two sub-networks, where at the first one, a convolutional neural network is used to treat the acoustic signal as an image, that captures spectral shape by operating on MFFC derived features and

TABLE I
REFERENCES ABOUT FEMH

References	Features	Classification method	Results
[3]	OpenSMILE 6552 acoustic features	BayesNet, Random Forest	Final score of 79.31%
[4]	MFCC	SVM, RF, KNN, GB, EL	Accur. SVM 0.6495, RF 0.6645, KNN 0.6603, GB 0.6735, EL 0.6848
[5]	MFCC+delta-delta, MFCC+spectral	NN, RF, SVM	Sensitivity NN SVM (MFCC+spectral) 96.9% 20.0%
[6]	NNE, Cepstral HNR, GNER MFCC	GMM, GBT	MEEI 99.4% SVD 93.2% Detection Score 72%
[7]	Transfer learning	SVM	Sensitivity 94.90%
[3]	Mel scaled spectrograms and MFCC	5-layer CNN and RNN	Sensitivity 96% Specificity 18%
[8]	MFCC 3rd-6th, SC, Spectral Flux, and ZCR	ATSVM	UAR of 60.67%.
[9]	JMFCC 3rd-6th, SC, Spectral Flux, ZCR	TSVM	UAR score of 60.67%
[10]	28 acoustic parameters	PCA auto-associative NN	86% - 100%
[11]	MFCCs, SWCE, multipeak, Thomson tapers	GM	Thomson multitaper outperforms functional, organic dysphonia

TABLE II
REFERENCES CONCERNING FEMH DATASET WITH DEMOGRAPHIC DATA INCLUDED

References	Features	Classification method	Results	Pathologies
[12]	Demogr., Symptom.	DT, LNA, KNN, SVXM, ANN	Demog. Neoplasm, PH, Symptoms VP	Neoplasm, PH, VP
[13]	MFCC, MFCC + delta	SVM, GMM, DNN	DNN outperforms	Nodule, Polyp., Neoplasm, Dysphonia, Sulcus
[2]	MFCC+delta, MFCC	GMM, DNN	Accuracy increase 2.02–10.32%	Neoplasm, PH ,VP

simple filterbank outputs and at the second one, a feed-forward network is analysing an enchanced input vector, consisting not only by the demographic parameters but with mid-term signal features as well.

C. NEURAL NETWORK ARCHITECTURE

We propose an architecture consisting of a combination of two sub-networks, where each sub-network takes over the processing of a different input source, as follows

- The first sub-network consists of four consecutively convolutional layers. Each layer contains 64, 64, 32 and 32 convolutional masks respectively, each one with a kernel of size 3×3. The output of each convolutional operation is processed through a ReLu function and the resulting feature matrix is subsequently subsampled by a max pooling layer with size 2×2. Each 52×64 input matrix produced by the preprocessing stage is passed through the convolutional layers.
- The second sub-network is a feed forward neural network with two hidden layers, consisting of 64 and 32 units, respectively, with Rectified linear unit activation functions. Each 37×1 input vector is fed to the input of the fully connected sub-network.
- Subsequently, the outputs of the aforementioned sub-networks are concatenated and fed to a dense layers with 1024 nodes, with each node's output being processed by a ReLu activation function. Finally, a softmax output layer of four units is used to produce posterior probability estimations of the four classes of the problem under study.

III. EXPERIMENTS

Given the number of design and training parameters that are usually involved in deep learning classifiers, a large number of trials was carried out regarding the number of layers, filter dimensions and activation functions of the adopted neural network architecture. Furthermore, the effects of crucial learning parameters, such as learning rate, backpropagation algorithm, size of training batches and dropout rate, were studied.

Concerning feature selection, various features and extraction methods were evaluated, using literature as a guide. For example, MFCCs combined with pitch frequency have been reported to produce a 99.44% classification accuracy for the binary problem of discriminating normal from pathological speech for the case of the sustained vowel /a/ [16]. Perturbation methods (including jitter and shimmer), signal-to-noise ratio and nonlinear dynamic methods (such as correlation dimension and second-order entropy) were used to analyze sustained and running vowels with laryngeal pathologies in [17]. Complexity measures of noise parameters and MFCCs were used in [18] and the wavelet packet transform on dysphonic voices was applied in [19]. Especially for dysphonia detection in recordings of the sustained vowel /a/, biologically inspired AM Analysis features [20], modulation spectral features combined with MFCCs [21] and modulation spectral features [22] have been reported.

Based on the above and taking into account the 2018 FEMH results, after experimentation, we chose to augment the standard MFCC vector with its first order derivative, and subsequently concatenate it with the logarithm of the mel-filterbank outputs, aiming to capture the spectral shape of the speech signal and its evolution over time. Furthermore, we used the 3 mid-term perturbation features that were described in the feature extraction section. All these features were eventually fused with metadata from medical records.

Concerning network design, we first experimented with a

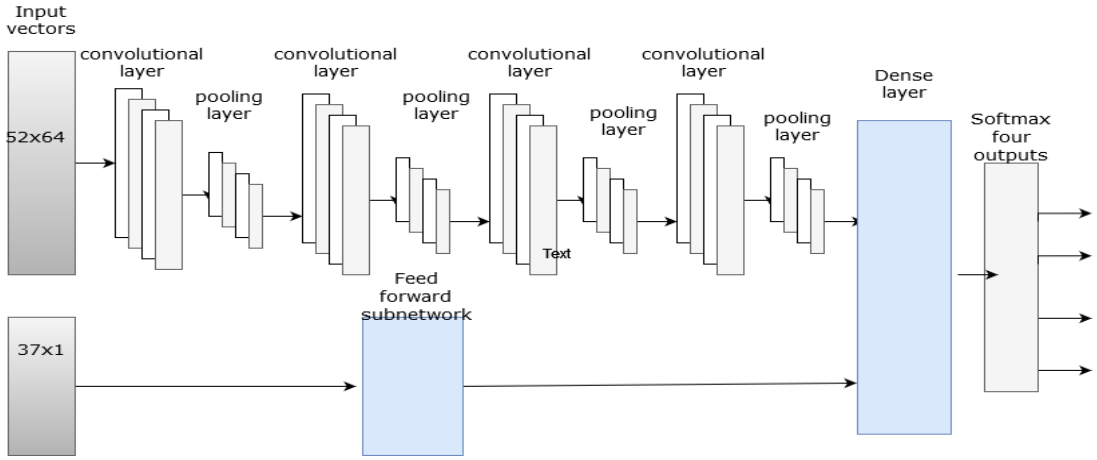

Fig. 1. Network architecture

combination of convolutional and recurrent neural network architectures as the two sub-networks of the scheme but this was eventually outperformed by a fusion of convolutional and feed forward architectures. A grid search on the number and size of network layers, activation functions and filter dimensions was carried out for the sake of network tuning. The proposed classifier was trained for 300 epochs using the Adam gradient descent algorithm to optimize the cross-entropy loss function, with a learning rate of 0.001, in a 5-fold cross-validation scheme. An early-stopping criterion was activated if the loss-value did not decrease significantly between epochs. Dropout regularization was used to reduce overfitting, with the dropout value set to 0.5 for the fully connected layer and 0.1 for each convolutional layer.

Of course, one of our main goals, was to test the influence of the integration of data from medical records into the system. To that end, four system setups were tested. At first, the convolutional sub-network was omitted, and only the perturbation features efficiency was tested, resulting into a classification accuracy of 46.5%. Then, the medical records were analysed alone, resulting into a classification accuracy of 58.5%. Then, with the combination of perturbation features and medical records, classification accuracy raised to 62.5%. The use of MFCCs in a convolutional setup, as a second information source, led to a further performance increase (65.9%). This is also shown in Table III where it can be seen that the integration of metadata from medical records with low-level signal descriptors and selected mid-level features yields the best feature combination.

The 2019 FEMH challenge required that the performance is reported with respect to classification accuracy. Our system had achieved a 65.9% classification accuracy on the public training set. This performance value dropped to 57% on the undisclosed testing set, giving our method the fifth place, based on official announced results. This performance drop can be perceived as some sort of overfitting, which we have tried to address in this paper, as we explain in the following section on critical issues.

TABLE III
RESULTS FOR DIFFERENT INPUT DATA COMBINATIONS

Input features	Results (Accuracy %)
Perturbation features	46.5 (mean over folds), {0.525 0.35 0.45 0.5 0.5} per fold
Medical Records	58.5 (mean over folds), {0.725 0.45 0.65 0.625 0.475} per fold
Medical Records and perturbation features	62.5 (mean over folds), {62.5 75 62.5 55 57.5} per fold
Medical Records Perturbation and MFCC	65.9 (mean over folds), {60 70 67.5 62.5 70} per fold

IV. AN OVERVIEW OF CRITICAL ISSUES

A main difficulty in this classification problem is the variability of the duration of the recordings and the fact that duration can be important for certain types of pathology. For example, in vocal palsy, the short duration of the recordings is indicative of the pathology. This is why we decided to segment recordings into mid-term segments of fixed duration ($\approx 1.3s$) and apply zero-padding to very short segments. By summing up the softmax outputs of the classifier for the various segments and selecting the maximum resulting value, the classification accuracy in our cross-fold validation setup increased to 71.5%. It is worth noting that the method that was submitted to the challenge was only selecting the first 5 seconds of each recording (or less depending on recording length) which yields lower performance on the experimental setup of this paper.

Furthermore, an important problem of the FEMH dataset is the small amount of training data, which can be a severe limitation when training deep learning architectures. In particular, the training dataset consists of only two hundred recordings for all four types of examined pathology. Previous work has shown that the introduction of noise to a network can, in some circumstances, lead to significant improvements in performance generalization [23]. Therefore, in order to artificially augment the available training dataset, noise injection is applied. Noise is added only at the input vectors and not on layer activation,

weights, gradients or outputs. The common approach of adding Gaussian noise is adopted [24] and noise is only added during the training phase. Of course, noise injection is not used during the evaluation of the model. This approach led to a slight improvement at the classification accuracy of the model, reaching 74.9%. More detailed results are shown in Table IV.

As a final remark, the results reported in this paper refer to a cross-fold validation scheme that is only based on the training data of the challenge. This was imperative because the testing dataset of the challenge, on which the ranking of methods was computed, has not been made publicly available by the organizers of the challenge up to the time of writing this paper.

TABLE IV
RESULTS FOR SEGMENTATION-AUGMENTATION TECHNIQUES

Segmentation	Augmentation	Results (Accuracy%)
Applied	Applied	74.99 (mean over folds), {75 80 72.5 70 75} per fold
Applied	Excluded	71.5 (mean over folds), {80 70 65 70 72.5} per fold
Excluded	Excluded	65.9 (mean over folds), {60 70 67.5 62.5 70} per fold

V. CONCLUSIONS

Our experimental study indicates that a carefully designed neural network architecture can serve as a promising classification method for the hard problem of voice pathology detection under the constraint of insufficient training data. In particular, we have shown that it is possible to treat MFCC-derived features and data from medical records as two different input sources to a single neural network architecture consisting of two sub-networks, thus implementing a competitive method in the context of the 2019 FEMH challenge. We also discussed how to deal with the variable length of speech recordings and investigated how data augmentation affects classification performance.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work has been partly supported by the University of Piraeus Research Center.

REFERENCES

[1] K. Daoudi and B. Bertrac, "On classification between normal and pathological voices using the meei-kaypentax database: Issues and consequences," in *INTERSPEECH-2014*, 2014.

[2] S.-H. F. et al., "Combining acoustic signals and medical records to improve pathological voice classification," 2019.

[3] C. Bhat and S. K. Kopparapu, "Femh voice data challenge: Voice disorder detection and classification using acoustic descriptors," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5233–5237, 2018.

[4] M. Pham, J. Lin, and Y. Zhang, "Diagnosing voice disorder with machine learning," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5263–5266, 2018.

[5] K. Degila, R. Errattahi, and A. E. Hannani, "The ucd system for the 2018 femh voice data challenge," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5242–5246, 2018.

[6] J. D. Arias-Londoño, J. A. G. García, L. Moro-Velázquez, and J. I. Godino-Llorente, "Byovoz automatic voice condition analysis system for the 2018 femh challenge," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5228–5232, 2018.

[7] K. A. Islam, D. Pérez, and J. Li, "A transfer learning approach for the 2018 femh voice data challenge," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5252–5257, 2018.

[8] M. Ju, Z. Jiang, Y. Chen, and S. Ray, "A multi-representation ensemble approach to classifying vocal diseases," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5258–5262, 2018.

[9] T. Grzywalski, A. Maciaszek, A. Biniakowski, J. Orwat, S. Drgas, M. Piecuch, R. Belluzzo, K. Joachimiak, D. Niemiec, J. Ptaszynski, and K. Szarzynski, "Parameterization of sequence of mfccs for dnn-based voice disorder detection," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5247–5251, 2018.

[10] Z.-Y. Chuang, X.-T. Yu, J.-Y. Chen, Y.-T. Hsu, Z. Xu, C.-T. Wang, F.-C. Lin, and S.-H. Fang, "Dnn-based approach to detect and classify pathological voice," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5238–5241, 2018.

[11] M. Pishgar, F. Karim, S. Majumdar, and H. Darabi, "Pathological voice classification using mel-cepstrum vectors and support vector machine," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5267–5271, 2018.

[12] S.-Y. Tsui, Y. Tsao, C.-W. Lin, S.-H. Fang, F.-C. Lin, and C.-T. Wang, "Demographic and symptomatic features of voice disorders and their potential application in classification using machine learning algorithms." *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics*, vol. 70 3-4, pp. 174–182, 2018.

[13] S.-H. F. et al., "Detection of pathological voice using cepstrum vectors: A deep learning approach." *Journal of voice : official journal of the Voice Foundation*, 2018.

[14] J. P. Teixeira and P. O. Fernandes, "Acoustic analysis of vocal dysphonia," *Procedia Computer Science*, vol. 64, pp. 466–473, 2015.

[15] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, 1993, pp. 97–110.

[16] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, vol. 1. IEEE, 2002, pp. 182–183.

[17] Y. Zhang and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *Journal of Voice*, vol. 22, no. 1, pp. 1–9, 2008.

[18] J. D. Arias-Londono, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.

[19] C. D. P. Crovato and A. Schuck, "The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 10, pp. 1898–1900, 2007.

[20] N. Malyska, T. F. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically-inspired amplitude-modulation features," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–873.

[21] M. Markaki, Y. Stylianou, J. D. Arias-Londoño, and J. I. Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5162–5165.

[22] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[23] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Computation*, vol. 8, pp. 643–674, 1996.

[24] L. Holmström and P. Koistinen, "Using additive noise in back-propagation training," *IEEE transactions on neural networks*, vol. 3 1, pp. 24–38, 1992.