Application of Neural Networks to Speech/Music/Noise Classification in Digital Hearing Aids

LORENA ÁLVAREZ
University of Alcalá
Polytechnic School
28805 Alcalá de Henares, Madrid
SPAIN
lorena.alvrezp@uah.es

COSME LLERENA
University of Alcalá
Polytechnic School
28805 Alcalá de Henares, Madrid
SPAIN
cosme.llerena@uah.es

ENRIQUE ALEXANDRE
University of Alcalá
Polytechnic School
28805 Alcalá de Henares, Madrid
SPAIN
enrique.alexandre@uah.es

Abstract: This paper focuses on the development of an automatic sound classifier embedded in a digital hearing aid aiming at enhancing the listening comprehension when the user goes from a sound environment to another different one. The approach we propose in this paper consists in using a neural network-(NN-) based sound classifier that aims to classify the input sound signal among speech, music or noise. The key reason that has compelled us to choose the NN-based approach is that neural networks are able to learn from appropriate training pattern sets, and properly classify other patterns that have never been found before. This ultimately leads to very good results in terms of higher percentage of correct classification when compared to those from other popular algorithms, such as, for instance, the k-nearest neighbor (k-NN) or mean square error (MSE) classifier, as clearly shown in the results obtained in this paper.

Key–Words: Neural networks, k-nearest neighbor, Linear classifier, Sound classification, Digital hearing aids

1 Introduction

A particular application that would be considered as very useful by most of hearing aid users, especially by the elderly, is that in which the hearing aid *itself* classifies the acoustic environment that surrounds its user and automatically selects the "amplification program" that is best adapted to such environment. This is referred to as "self-adaptation". The "manual" approach, in which the user has to identify the acoustic environments where he or she is in and then choose the adequate program that best fits this situation by using a switch on the hearing instrument or some kind of remote control, is very uncomfortable and frequently exceeds the abilities of many hearing aid users [1]. Additionally, even for normal hearing people, it is not always clear which program should be selected for best performance and considering that elderly people are the main group of hearing aid users, it cannot be expected that they will handle this task properly.

Therefore, it seems to be apparent from the previous paragraph the necessity for hearing aids can be automatically fitted according to the preferences of the user for various listening conditions. This type of hearing aid would help the user in improving the speech intelligibility, increasing his or her comfort level and allowing him or her to lead a normal life [2]. As a matter of fact, in a study with hearing-impaired people, it was observed that the automatic switching mode of the instrument was deemed useful by a majority of test subjects, even if its performance was not always perfect [3].

Although the necessity of the aforementioned sound classification system seems to be clear, its implementation is, on the contrary, a difficult matter. Despite the impressive advances in microelectronics, the development of an automatic sound classification system in a digital hearing aid is a challenging goal. The underlying reason is that most of hearing aids in the market have very strong constraints in terms of computational capacity, memory and battery, which, for the sake of clarity, we will summarize in the section that follows. These design restrictions severely limit the implementation of complex algorithms on a digital hearing aid.

Regarding the mentioned issues, the purpose of this work is just the development of a sound classifier, which programmed on a hearing aid based on a digital signal processor (DSP), assists it to enhance the user's listening skills. For this goal to be reached, we have explored the feasibility of some kind of neural network (NN) tailored for running on a hearing aid and able to classify the input sound signal among speech, music or noise. The reason for which these three classes have been selected, in detriment of an approach with more classes, is because we have considered these acoustic environments to be as the most

ISBN: 978-1-61804-027-5

comfortable for facing the everyday life for most of hearing aid users.

In the effort of making the paper to stand by self, after summarizing the important limitations t system suffers from (Section 2), Section 3 will intr duce the implemented classification system, descri ing the input features (Section 3.1) and the classifi used (Section 3.2). Although the main classifying a gorithm is based on neural networks, the results o tained will be compared with those achieved by tv algorithms proposed in the literature: the k-neare neighbor (k-NN) and the mean square error (MSI In the effort of clearly understanding how these tv algorithms work, they will be also described in Se tion 3.2. The paper is completed with the experime tal work (Section 4) and the discussion of the resu (Section 5).

2 A Brief Overview of the Hardwa Limitations of Hearing Aids

As mentioned in the Introduction, digital hearing ai suffer from important constraints and, in this respe Figure 1 will assist us in introducing the key concep that strongly affect the design of our NN-based clas fier.

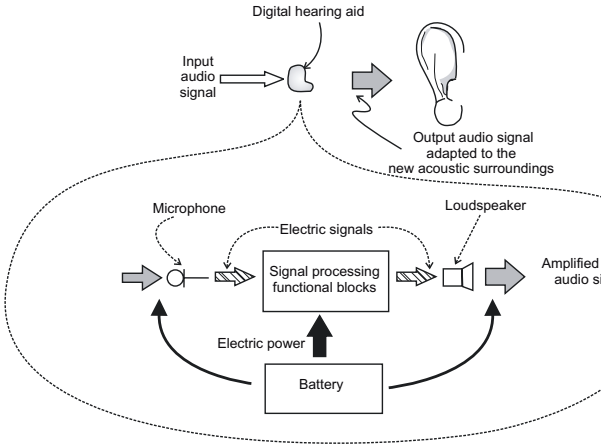

Figure 1: Simplified diagram of the structure of a typical digital hearing aid.

A typical digital hearing aid basically consists of:

- A microphone to convert the input sound into an electrical signal.
- A number of electronic blocks aiming at compensating the hearing losses the user suffers from.
- A tiny loudspeaker to convert the electrical signal back to sound.
- A small battery to supply electric power to the electronic devices that compose the hearing instrument.

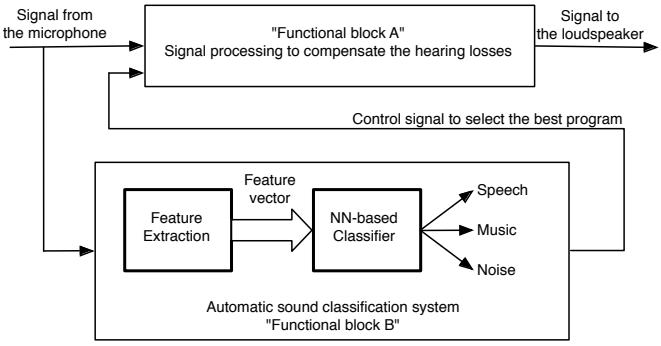

Figure 2: Conceptual representation of the functional blocks to be implemented in the hearing aid.

Note that the aforementioned classifying system consists conceptually of two basic parts:

- A feature extraction stage in order to properly characterize the input sound signal to be classified.
- A NN-based, three classes (speech/music/noise) classifier.

The aforementioned restrictions arise mainly from the tiny size of the hearing aid, specially for the in-the-canal (ITC) and completely-in-the-canal (CIC) models, which, as illustrated in Figure 1, must contain not only the electronic equipment, but also the battery. The DSP -which integrates the CPU core, the A/D and D/A converters, the filter-bank, the RAM, ROM, and EPROM memories, and some input/output ports- has to work at a very low clock frequency in order to minimize power consumption and thus maximize the battery life. Note that the power consumption must be low enough to ensure that neither the size of the battery pack nor the frequency of battery changes will annoy the user.

With this in mind, it seems to be clear that the tailoring of the NN requires a *balance* consisting in

reducing the computational demands, without degrading the performance perceived by the user.

3 The Proposed System

As stated beforehand, it basically consists of a feature extraction block, and the aforementioned classifier based on a neural network.

3.1 Feature Extraction

There is a number of interesting features that could potentially exhibit different behavior for speech, music and noise, and thus may help the system properly classify the input sound signal. In the effort of carrying out the experiments of this paper, we have selected a subset of them that are very well-known in hearing aids and can be computed on them with low computational cost [4].

These features will be now briefly described, although more detailed descriptions can be found in [5, 6, 7].

- Spectral centroid: this parameter can be associated with the measure of the timbre of a sound, and, from a mathematical viewpoint, it is obtained by evaluating the center of gravity of the spectrum:

$$C_i = \frac{\sum_{k=1}^{N_B} |\chi_i(k)| \cdot k}{\sum_{k=1}^{N_B} |\chi_i(k)|} \tag{1}$$

  where $\chi_i(k)$ represents the k-th frequency bin of the spectrum at frame i and N_B is the number of frequency bands available in the DSP ($N_B = 64$, in our case).

- Voice2white: it is a measure of the energy inside the typical speech band (300-4000 Hz) when compared to the whole energy of the signal $x_i(t)$. Mathematically,

$$V_i = \frac{\sum_{k=M_1}^{M_2} |\chi_i(k)|}{\sum_{k=1}^{N_B} |\chi_i(k)|} \tag{2}$$

  M_1 and M_2 being the limits of speech spectrum.

- Spectral flux: this feature is defined as the average variation value of spectrum between two adjacent frames:

$$F_i = \sum_{k=1}^{N_B} (|\chi_i(k)| - |\chi_{i-1}(k)|)^2. \tag{3}$$

  It is associated with the amount of spectral local changes between two consecutive frames.

- Short time energy: this simple measure is defined as the mean energy of the signal within each analysis frame. It is computed using the expression:

$$E_i = \frac{1}{N_B} \sum_{k=1}^{N_B} |\chi_i(k)|^2. \tag{4}$$

Finally, these features are calculated by estimating the mean value and the standard deviation of these measurements for M different time frames. Therefore, the feature extraction algorithm generates the following feature vector: $\mathbf{F} = [\widehat{E}[\mathbf{F}_1], \widehat{\sigma}^2[\mathbf{F}_1], \ldots, \widehat{E}[\mathbf{F}_{n_f}], \widehat{\sigma}^2[\mathbf{F}_{n_f}]]$ ($n_f = 4$, in our case), its dimension being $\dim(\mathbf{F}) = 2 \cdot n_f = L$. This is just the signal-describing vector that feeds the classifier. For the sake of clarity, it is written formally $\mathbf{F} = [F_1, \ldots, F_L]$.

3.2 Classifier

As previously mentioned, the main classification algorithm considered in this work is a NN-based classifier. The results achieved by this classifier will be compared with those obtained when using the mean square error (MSE) or k-nearest neighbor (k-NN) classifier. These algorithms will be now briefly described, along with a brief description of neural networks.

3.2.1 Mean square error (MSE) classifier

Mean square error (MSE) classifier has been explored here because of its simplicity and good results. In this kind of linear classifier, the decision rule depends on a linear combination of the input features that have been computed in the feature extraction stage:

$$y = b + \sum_{n=1}^{L} x_n w_n \tag{5}$$

where x_n represents the values of the n-th feature, L represents the number of input features, b the bias value, and w_n the weights of the linear combination. In order to obtain a decision, C different evaluations of the expression above are calculated, one for each class. The final decision corresponds to the linear combination with the highest result.

This process can be described using matrix notation. Let us define the input patterns matrix as:

ISBN: 978-1-61804-027-5

$$\mathbf{Q} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{L1} & x_{L2} & x_{L3} & \dots & x_{LN} \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} \quad (6)$$

where N represents the number of input patterns, and L the dimension of each pattern. Note that the last row equals 1 in order to define the weights of the c as:

$$\mathbf{V} = \begin{pmatrix} w_{11} & w_{21} & \dots & w_{L1} & b_1 \\ w_{12} & w_{22} & \dots & w_{L2} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{1C} & w_{2C} & \dots & w_{LC} & b_C \end{pmatrix}$$

where C represents the number of classes to ($C = 3$, in our case).

The output of the classifier can be definec

$$\mathbf{Y} = \mathbf{V} \cdot \mathbf{Q}$$

Y being a matrix with C rows and N column

The error can be defined as:

$$\mathbf{E} = \mathbf{Y} - \mathbf{T} = \mathbf{V} \cdot \mathbf{Q} - \mathbf{T}$$

where T represents a $C \times N$ matrix contai target classes for each input pattern. If we de mean square error as:

$$MSE = \frac{1}{NC} \sum_{n=1}^{N} \sum_{c=1}^{C} e_{cn}^2$$

we can therefore derive with respect to the coe w_{ij} and minimize the MSE. The result obt found to be:

$$\mathbf{V} = \mathbf{T} \cdot \mathbf{Q}^T \cdot \left(\mathbf{Q} \cdot \mathbf{Q}^T\right)^{-1}.$$

3.2.2 k-Nearest Neighbor

The k-nearest neighbor (k-NN) is also a very yet powerful classification algorithm. To bette stand it, let us assume that we have a training N vectors grouped into C different classes. Pu simple, to obtain the class corresponding to a served vector $\mathbf{x}$, the algorithm has simply to the k nearest neighbors to the test vector $\mathbf{x}$, an their class numbers they belong to, usually majority rule. Although it is possible to us ent distance measures, most implementations an euclidean measure.

3.2.3 Neural networks

Figure 3 shows the basic architecture of a NN consisting of three layers (input, hidden and output layers) interconnected by links with *adjustable weights* [8]. This figure also depicts the NN activation function considered as more appropriate for the problem at hand: the logarithmic sigmoid for hidden and output neurons.

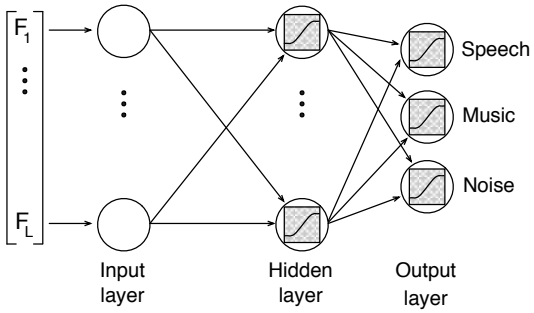

Figure 3: Neural network consisting of three layers: input, hidden and output.

In this NN architecture, the number of input neurons corresponds to that of the features used to characterize the sound (that is, the dimension of feature vector), the number of output neurons is related to the three classes we are interested in, and finally, the number of hidden neurons depends on the adjustment of the complexity of the network [8]. If too many hidden neurons are used, the capability to generalize will be poor; on the contrary, if too few hidden neurons are considered, the training data cannot be learned satisfactorily. In this respect, selecting the appropriate number of hidden neurons becomes a fundamental issue when designing neural networks.

4 Results

Prior to the description of the batches of experiments carried out and the discussion of the corresponding results (Section 4.2), it is convenient to describe the sound database used in the experiments (Section 4.1).

4.1 Sound database

It consists of a total of 7299 audio files, with a length of 2.5 seconds each. These files include speech-in-quiet, speech-in-noise, speech-in-music, vocal music, instrumental music and noise. This database has been manually labeled, obtaining a total of 807 files of speech-in-quiet, 1615 files of both speech-in-noise

and speech-in-music, 2440 files of both vocal and instrumental music and, finally, 2437 files of noise. All audio files are monophonic and were sampled with a sampling frequency of 16000 Hz with 16 bits per sample. We have taken into account a variety of noise sources, including those from the following diverse environments. For the sake of generality, speech files were recorded from different sources, with varying grades of reverberation. The files that correspond to speech-in-noise that we explored exhibit different signal to noise ratios (SNRs) ranging from 0 to 10 dB. In a number of experiments, these values have been found to be representative enough of the following fact related to perceptual criteria: lower SNRs could be treated by the hearing aid as noise, and higher SNRs could be considered as clean speech.

Each sound file in the database has been filtered using the hearing aid simulator described in [9] without feedback.

For training, validation and testing, it is necessary for the database to be divided into three different sets. These sets include 2905 files ($\approx$ 36%) for training, 985 files ($\approx 14\%$) for validation, and 3409 files (50%) for testing. This division was done randomly, ensuring that the relative proportion of files of each category is preserved for each set.

The results we illustrate below show the percentage of correct classification for the test set.

4.2 Results obtained

Table 1 shows the mean percentage of correct classification (%) obtained in the 20 runs of the training process for a variety of multilayer perceptrons (MLPs) with different numbers of hidden neurons, when using the mean and variance of features listed in Section 3.1. The Levenberg-Marquardt algorithm [8, 10] with Bayesian regularization [8] has been found to be the most appropriate method for training the neural network. The main advantage of using regularization techniques relies on the fact that the generalization capabilities of the NN ends in being improved, making it capable of reaching good results even with smaller networks, since the regularization algorithm itself prunes those neurons that are not strictly necessary.

Turning again our attention to Table 1, MLP W labels the corresponding NN is a multilayer perceptron with W neurons in the hidden layer. The batches of experiments have explored a number of hidden neurons ranging from 1 to 20. A higher number of hidden neurons has been found to be unfeasible because of the greater associated computational cost.

As illustrated, the best result is obtained with MLP 19, or in other words, 19 hidden neurons, with a

Classifier	Percentage of correct classification (%)
MLP 1	70.3
MLP 2	73.4
MLP 3	77.8
MLP 4	79.7
MLP 5	80.0
MLP 6	80.1
MLP 7	80.8
MLP 8	80.7
MLP 9	80.6
MLP 10	81.3
MLP 11	81.6
MLP 12	81.2
MLP 13	80.7
MLP 14	81.4
MLP 15	80.9
MLP 16	81.0
MLP 17	80.7
MLP 18	80.8
MLP 19	81.9
MLP 20	81.3

Table 1: Mean percentage of correct classification (%) of different classifiers for speech/music/noise classification. MLP W means multi-layer perceptron with W neurons in the hidden layer.

percentage of correct classification $P_{\text{CC}} \approx 82.0$ %.

In the aim of clearly quantifying to what extent the use of neural networks, for automatic sound classification in hearing aids, is effective or not, we have explored the feasibility of using, for the classification problem at hand, two different classification algorithms proposed in the literature: the k-nearest neighbor (k-NN) algorithm and the mean square error (MSE) classifier. The results obtained, when using these two algorithms, are illustrated in Table 2.

It is interesting to note that the value of k of the k-NN algorithm is an user-specific parameter. In many articles, it is automatically selected in order to maximize the percentage of correct classification over the validation set. Just in this respect, different k-NN classifiers with values of k from 1 to 20 have been implemented in this work, and the value of k that achieves the best correct classification rate over the validation set has been selected. This value has been found to be $k = 5$ in our experimental work.

As shown, these both classifiers perform worse

ISBN: 978-1-61804-027-5

Classifier	Percentage of correct classification (%)
MSE	70.6
5-NN	76.9

Table 2: Mean percentage of correct classification (%) of mean square error (MMSE) and k-nearest neighbor (k-NN) classifiers for speech/music/noise classification.

that the novel strategy: the NN-based classifier has a mean percentage of correct classification ($P_{\text{CC}} = 81.9\%$), which is higher than any of those achieved by MSE or k-NN classifier ($P_{\text{CC}} = 70.6\%$ or $P_{\text{CC}} = 76.9\%$, respectively).

Therefore, these results are very promising in the sense that the application of neural networks to the task of speech/music/noise classification in hearing aids seems to provide very good results.

5 Discussion

In this paper we have explored a three-classes neural network-based classifier to be implemented in hearing aids in the effort of solving their irregular use. Although hearing losses disqualify many people from holding a normal life, however, many of them do not make use of hearing aids. This is because many hearing aids in the market cannot automatically adapt to the changing acoustical environment the user daily faces on. Within this framework, this paper has focused on the development of the mentioned automatic sound classifier for digital hearing aids that, constrained to the computational limitations of these devices, aims to enhance the listening comprehension when the user goes from a sound environment to another different one.

Just in this respect, we have implemented a neural network-based classifier that aims to classify the input sound signal among speech, music or noise. In order to check the results, we have carried out a number of experiments to compare the results obtained when using neural networks with those achieved by two classical algorithms proposed in the literature: k-nearest neighbor (k-NN) or mean square error (MSE). This comparison has been evaluated in terms of performance. The experiments prove that both k-NN and MSE classifiers perform worse than the NN-based classifier. To illustrate this, it is worth mentioning that the k-NN obtains a percentage of correct classification equal to 76.9%, while the NN-based classifier increases the percentage of correct classification up to 81.9%. Please note that we only mention the result corresponding to k-NN algorithm and not that corresponding to MSE classifier, because this latter classifier exhibits worse performance than k-NN algorithm.

Acknowledgements: The work has been partially funded by the Spanish Ministry of Science and Innovation, under project TEC2009-14414-C03-03, and by the Community of Madrid under project CCG10-UAH-TIC-5907.

References:

[1] M. C. Büchler, S. Allegro, S. Launer, and N. Dillier. "Sound classification in hearing aids inspired by auditory scene analysis", *EURASIP Journal on Applied Signal Processing* vol. 2005, no. 18, pp. 2991-3002, 2005.

[2] V. Harnacher, J. Chalupper, J. Eggers, et al. "Signal processing in high-end hearing aids: state of the art, challenges, and future trends", *EURASIP Journal on Applied Signal Processing* vol. 2005, no. 18, pp. 2915-2929, 2005.

[3] M. C. Büchler, *Algorithms for sound classification in hearing instruments*, Ph.D. thesis, Swiss Federal Institute of Technology, Zurich, Switzerland, 2002.

[4] L. Cuadra, R. Gil-Pita, E. Alexandre, and M. Rosa-Zurera, Joint design of Gaussianized spectrum-based features and least-square linear classifier for automatic acoustic environment classification in hearing aids. *Signal Processing* 90, 8 (August 2010), 2628-2632.

[5] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", in *ICASSP*, 1997.

[6] E. Guaus, E. Battle, "A non-linear rhythm-based style classification for broadcast speech-music classification", in *AES 116th Convention*, 2004.

[7] L. Lu, H. J. Zhang, H. Jiang, "Content analysis for audio classification and segmentation", in *IEEE Transactions on Speech and Audio Proceesing*, 10(7) pp. 504-516, 2002.

[8] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2001.

[9] R. Vicen-Bueno, R. Gil-Pita, M. Utrilla-Manso, L. Álvarez-Pérez, "A hearing aid simulator to test adaptative signal processing algorithms".Proceedings of the *IEEE International Symposium on Intelligent Signal Processing (WISP)*, pp.619-624, 2007.

[10] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, UK, 1995.