EXPLORING THE FEASIBILITY OF A TWO-LAYER NN-BASED SOUND CLASSIFIER FOR HEARING AIDS

Enrique Alexandre, Lucas Cuadra, Lorena Álvarez, Manuel Utrilla

Dept. of Signal Theory and Communications
University of Alcalá, 28805, Alcalá de Henares, Spain
phone: + (34) 918856727, fax: + (34) 918856699, email: enrique.alexandre@.uah.es

ABSTRACT

This paper focuses on the development of an automatic sound classifier for digital hearing aids that aims to enhance the listening comprehension when the user goes from a sound environment to another different one. The approach consists in dividing the classifying algorithm into two layers that make use of two neural network algorithms that work more efficiently: the input signal discriminated by the first layer into either speech or non-speech is ulteriorly classified more specifically depending on whether the user is in a conversation (both in quiet and in the presence of background noise) or in a noisy ambient in the absent of speech. The system results in having three classes, labeled "speech in quiet", "speech in noise", and "noise". A brief discussion on the computational complexity of this approach illustrates its feasibility to be implemented on a conventional digital hearing aid.

1. INTRODUCTION

Hearing aids are usually designed and programmed for only one listening environment. However it has been shown that their users usually prefer to have different amplification schemes in different listening conditions [1][2]. Thus, modern digital hearing aids generally allow the user to manually select among different programs (different frequency responses or other processing options such as compression methods, directional microphone, feedback canceller, etc.) depending on the listening conditions. The user has therefore to recognize the acoustic environment and choose the program that best fits this situation by using a switch on the hearing instrument or some kind of remote control.

This indicates the need for hearing aids that can be automatically fitted according to user preferences in a variety of listening conditions. In a study with hearing-impaired subjects, 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 perfect [3].

The two most important listening environments for a hearing aid user are speech in quiet and speech in noise [4]. While the first situation is usually easy to handle, nevertheless speech in noise is a much more difficult environment for the hearing aid user as a consequence of its low signal-to-noise ratio. Therefore, automatic detection of noise in the listening environment can be helpful to the user, since it would allow switching on or off different features of the hearing aid, such as directional microphone or a noise suppression algorithm. In this respect, implementing such a classifier on the hardware resources of a hearing aid is a challenging goal. To understand this, it is convenient to have a look at the block structure of a digital hearing aid represented in Fig. 1. This consists basically of a microphone to convert sound into electric signal, a digital signal processing (DSP) integrated circuit (IC) and a small loudspeaker to convert the electric signal back to sound. Fig. 1 shows the two main functional blocks that must be implemented on the mentioned DSP. The first one, labeled "functional block A", corresponds to the set of signal processing stages aiming to compensate the hearing losses. The second one ("functional block B") is the classifying system itself. The key point is that the DSPs used in hearing aids have generally very considerable constraints in terms of computational capacity and memory, which must be taken into account when implementing the classifier.

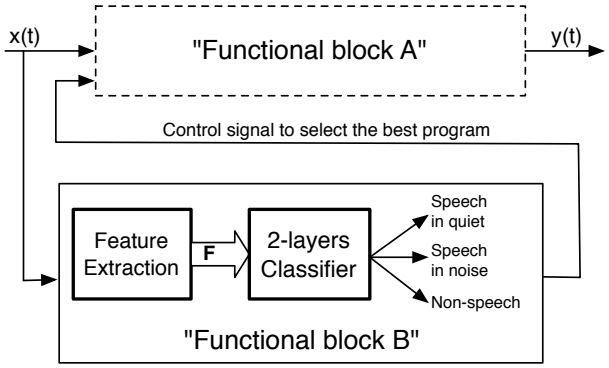

Figure 1: Simplified block diagram of a DSP-based hearing aid. **F** labels the feature vector describing the signal, $x(t)$ the input signal from the microphone and $y(t)$ the output signal to feed the loudspeaker.

The purpose of this work is the development of a two-layer, NN-based sound classifier, which programmed on a DSP-based hearing aid, assists it to enhance the user's listening skills. As it has been commented, the distinction between speech and any other signal is the crucial task. This is just the reason that compels us to explore a divide-and-conquer strategy that leads to a classification systems composed of the two layers represented in Fig. 2. The first one discriminates the input sound into either speech or non-speech, this second category being named noise in our work, because particular emphasis is put on speech intelligibility. If the discriminated signal has been found to be speech, a second algorithm in the

This work has been partially financed by the Universidad de Alcalá (UAH PI2005/081), Comunidad de Madrid/Universidad de Alcalá (CAM-UAH2005/036, CCG06-UAH/TIC-0378) and the Spanish Ministry of Education and Science (TEC2006-13883-C04-04/TCM).

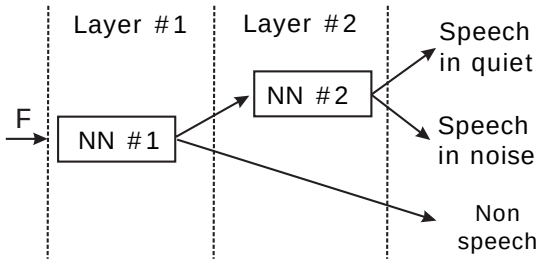

Figure 2: Scheme of the proposed system, consisting of two NN-based classifiers arranged in a two-layer structure.

second layer classifies it into either speech in quiet or speech in noise. The hearing aid will thus select the listening program best adapted to the current environment by means of the control signal illustrated in Fig. 1.

The paper is structured as follows: first the implemented method will be described, including the feature extraction process, the classification algorithm and the sound database used for the experiments. After that, some results will be shown, to illustrate the behavior of the proposed system. A brief discussion on the computational complexity of the system and the feasibility of its implementation on a hearing aid DSP will also be made. The paper will conclude with a discussion on the results obtained.

2. SYSTEM OVERVIEW

As represented in the "Functional Block B" in Fig. 1, the system we propose is composed of two basic stages: a feature extraction process, and the classifier itself.

2.1 Features Extraction

The input audio signal is divided into frames with a length of 64 samples, and with no overlap between adjacent frames. Then, a WOLA (Weighted Overlap-Add) filter bank with 64 bands is computed, and all the considered features are calculated. Finally, the mean and standard deviation values are estimated for a number of frames in order to soften the values.

The features that will be considered in this work, which have been chosen among the most popular in several audio classification applications, will be now briefly described. More detailed descriptions of these features can be found, for instance, in [5], [6] and [7].

2.1.1 Spectral Centroid.

The spectral centroid can be associated with the measure of brightness of a sound, and is obtained by evaluating the center of gravity of the spectrum:

$$Centroid_t = \frac{\sum_{k=1}^{N} |X_t[k]| \cdot k}{\sum_{k=1}^{N} |X_t[k]|} \qquad (1)$$

where $X_t[k]$ represents the k-th frequency bin of the spectrum at frame t, and N is the number of samples.

2.1.2 Spectral Roll-off.

The spectral roll-off ($RollOff_t$) is usually defined as the frequency bin below which a PR% of the magnitude distribution is concentrated:

$$\sum_{k=1}^{RollOff_t} |X_t[k]| = PR \cdot \sum_{k=1}^{N} |X_t[k]| \ . \qquad (2)$$

A typical value for PR is PR=85%. The spectral roll-off can give an idea of the shape of the spectrum.

2.1.3 Spectral Flux.

It is associated with the amount of spectral local changes, and is defined as follows:

$$Flux_t = \sum_{k=1}^{N} \left(|X_t[k]| - |X_{t-1}[k]|\right)^2 \ . \qquad (3)$$

2.1.4 Zero Crossing Rate (ZCR).

The ZCR is computed from the temporal signal $x[n]$ using the expression:

$$ZCR_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \qquad (4)$$

where $sign(\cdot)$ represents the sign function, which returns 1 for positive arguments and -1 for negative ones. This parameter takes higher values for noise and unvoiced speech than for voiced speech.

2.1.5 High Zero Crossing Rate Ratio (HZCRR).

This feature, proposed in [6], is computed from the ZCR, and is defined as the number of frames whose ZCR is 1.5 times above the mean ZCR on a window containing M frames.

It can be demonstrated [6] that the HZCRR takes higher values for speech than for music since speech is usually composed by alternating voiced and unvoiced fragments, while music does not follow this structure.

2.1.6 Short Time Energy (STE).

It is defined as the mean energy of the signal within each analysis frame.

2.1.7 Low Short-Time Energy Ratio (LSTER).

Similarly to the HZCRR, the LSTER is obtained from the STE, and defined as the ratio of frames whose STE is 0.5 times below the mean STE on a window that contains M frames.

2.1.8 Mel-Frequency Cepstral Coefficients (MFCCs).

These are a set of perceptual parameters calculated from the STFT [8] that have been widely used in speech recognition. They provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first coefficients. The application of these parameters for music modeling was discussed by Logan in [9]. To represent speech, 13 coefficients are commonly used, although it has been demonstrated that for classification tasks, it is enough to take into account only the first five coefficients [10].

2.1.9 Voice2White (V2W).

This parameter, proposed in [5], is a measure of the energy inside the typical speech band (300-4000 Hz) respect to the whole energy of the signal.

2.1.10 *Percentage of Low Energy Frames (LEF).*

It is defined as the proportion of frames with RMS power less than 50% of the mean RMS power within a one-second window [11].

2.1.11 *Loudness.*

Defined as an exponential function of the energy of the audio signal: $Loudness_t = Energy_t^{0.23}$.

2.1.12 *Spectral Flatness Measure (SFM).*

This feature gives an idea of the flatness of the spectrum, and according to [12], is defined as the relation between the geometric and arithmetic means of the power spectral density for each critical band.

2.1.13 *Bandwidth*

This feature is calculated from the spectral centroid:

$$BW = \frac{\sum_{k=1}^{N}(k - Centroid)^2 \cdot |X[k]|^2}{\sum_{k=1}^{N}|X[k]|^2} \quad (5)$$

2.2 Classification System

As it was commented before, the objective of this work is to classify the input audio signal as speech in quiet, speech in noise, or noise. While it would be possible to use a single classifier to distinguish among the three considered classes, this approach has some disadvantages for this particular application. To explain this, let us consider that the input signal is speech in quiet. If the classification algorithm confuses it with noise, the hearing aid will reduce the gain and the user will probably loose all the information. On the contrary, if the speech in quiet is confused with speech in noise, the hearing aid will switch on some mechanisms, in this case unnecessary, to reduce the noise, without affecting too much to the received speech information. From this it can be observed that the distinction between speech (with and without noise) and noise is much more critical in terms of maximum allowed probability of error than the distinction between speech in noise and speech in quiet.

As previously mentioned, to solve this problem a divide and conquer strategy was applied, that is, rather than using one single classifier, the use of two more specialized binary classifiers is proposed. Each one of these classifiers will be based on a neural network, and will be separately trained.

Neural networks can be viewed as massively parallel computing systems consisting of a large number of simple processors with many interconnections [13][14]. A three-layer feedforward backpropagation neural network (also called multilayer perceptron or MLP) was implemented.

The nodes in the hidden layer used a logsig activation function, while a linear transfer function was used for the nodes in the output layer. The weights of each node were adjusted using a gradient descent algorithm to minimize the mean squared error (MSE) between the output of the network for a certain training data set and the desired output. The network was trained using the Levenberg-Marquardt backpropagation algorithm [15] with bayesian regularization [16].

2.3 Network size considerations

One argument against the feasibility of neural networks for being used on DSP-based hearing aids consists in the, a priori, high computational complexity, which, among other topics, is related to the network size. However it is worth exploring its implementation because, as pointed out in [3] and [17], neural networks are able to achieve very good results in terms of probability of error when compared to other popular algorithms such as a rule-based classifier, the Fisher linear discriminant, the minimum distance classifier, the k-Nearest Neighbor algorithm, or a Bayes classifier.

The "negative" facet, as mentioned, could arise from the fact that the computational complexity of a neural network is the highest of all those classifiers. This complexity depends on the number of weights that need to be adapted, and consequently on the number of neurons which compose the neural network. In particular, the number of simple operations required by a neural network to produce one output is given by:

$$N_{op} = W(2L + 2M + 1) + 2M - 1 \quad (6)$$

where W, L and M are the number of hidden, input and output neurons respectively. Note that L equals the dimension of the feature vector.

From this, it can be observed that one way to achieve this goal is, as commented before, to decrease the number of input features (that is the dimensionality of the feature vector inputting the network), and thus the number of input neurons. It is for this reason that only one feature will be considered.

On the other hand, it will be also necessary to reduce the number of neurons in the hidden layer. As a rule of thumb, a number of hidden neurons equal to the logarithm of the number of training patterns has been empirically shown to be appropriate [18]. In our experiments, the number of hidden neurons was changed from 1 to 10 and the best value in terms of validation error was chosen.

2.4 Database Used

The sound database used for the experiments consisted of a total of 2936 files, with a length of 2.5 seconds each. The sampling frequency was 22050 Hz with 16 bits per sample. The files corresponded to the following categories: speech in quiet (509 files), speech in stationary noise (727 files), speech in non-stationary noise (728 files), stationary noise (486 files) and non-stationary noise (486 files). Noise sources were varied, including those corresponding to the following environments: aircraft, bus, cafe, car, kindergarden, living room, nature, school, shop, sports, traffic, train, train station. Music files, both vocal and instrumental, were also considered as noise sources. The files with speech in noise presented different Signal to Noise Ratios (SNRs) ranging from 0 to 10 dB.

The database was then divided into three different sets for training, validation and test, including 1074 (35%), 405 (15%) and 1457 (50%) files respectively. The division was made randomly and ensuring that the relative proportion of files of each category was preserved for each set.

3. RESULTS

This section presents the results obtained with the proposed system. For the sake of clarity, the results for each one of the classification tasks will be shown separately.

	Speech/Non-Speech	Clean/Noisy
Centroid	86.7% ($W=3$)	75.0% ($W=10$)
Roll-off	85.3% ($W=8$)	74.3% ($W=10$)
Spectral flux	76.4% ($W=6$)	76.3% ($W=9$)
ZCR	86.1% ($W=7$)	74.2% ($W=5$)
HZCRR	74.0% ($W=4$)	59.4% ($W=7$)
STE	80.5% ($W=9$)	79.1% ($W=7$)
LSTER	76.9% ($W=8$)	79.3% ($W=3$)
MFCC	87.4% ($W=4$)	84.7% ($W=4$)
V2W	84.9% ($W=3$)	80.3% ($W=8$)
LEF	66.3% ($W=6$)	60.7% ($W=8$)
Loudness	81.1% ($W=8$)	88.4% ($W=2$)
SFM	86.6% ($W=3$)	78.3% ($W=2$)
Bandwith	86.7% ($W=7$)	69.9% ($W=8$)
Best 5	92.2% ($W=5$)	95.0% ($W=6$)
All	96.6% ($W=24$)	95.9% ($W=14$)

Table 1: Probabilities of correct classification obtained for the speech/non-speech and clean/noisy speech tasks. The number of hidden neurons (W) is also indicated.

3.1 Speech/Non-speech Classification

The objective of this first task is to classify the input file as either speech or non-speech. Speech files include those with speech in quiet as well as those with speech in noise. Non-speech files are those with either music or background noise.

Table 1 shows the results obtained by the different algorithms and sets of features used. A MLP with three layers was trained for different numbers of hidden neurons. The experiment was repeated 10 times, and the best network in terms of validation error was selected. The results show the probability of correct classification achieved for the test set, jointly with the number of hidden neurons, W. The results obtained for the combination of the best 5 features and for all the features are also shown for comparative purposes.

As it can be observed, the best result is obtained with the Mel-Cepstrum Cepstral Coefficients (87.4%), followed by the spectral centroid (86.7%).

3.2 Clean/Noisy Speech Classification

The goal of this second task is to distinguish between speech in quiet and speech in noise. Table 1 shows the results obtained. As occurred in the Speech/non-speech classification task in Sect. 3.1, the shown probability corresponds to the network that exhibits the lower MSE for the validation set.

The best result is now achieved by the loudness feature, with a probability of correct classification equal to 88.4%, with only two neurons in the hidden layer.

4. CONSIDERATIONS ON REAL TIME IMPLEMENTATION

As it was stated before, our goal is the implementation of this classifier on a hearing aid. In our system, a sampling frequency of 22050 Hz is considered, with a frame length of 64 samples, which corresponds to $2.9\mu s$. This implies a rate of 344 frames per second. Since our DSP has a total computational power of approximately 3 MIPS, around 2 MIPS being already used, only 1 MIPS is thus available for the classification algorithm. This means that roughly speaking, a total number of 2900 instructions is available per frame.

From the previous section, the best feature for the speech/non-speech classification task is the set of MFCCs, which however, suffer from an excessively high computational complexity. As a consequence, it has been decided to use the spectral centroid, since it provides a similar result in terms of probability of correct classification with a much lower computational complexity. For the clean/noisy speech classification task it was decided to use the loudness, since its complexity is affordable in our system.

Note that both the total energy (needed to compute the spectral centroid) and the loudness value must be computed for other processing tasks included in Functional Block A, shown in Fig. 1. Their computational cost will therefore not be considered here.

With regards to the spectral centroid, the number of operations required has been found to be 64 products, 63 sums and 1 division. This makes a total of 128 simple operations. Nevertheless, to complete the calculation of the selected features it is necessary also to compute their mean and variance. For doing so the following estimators are considered:

$$\hat{\mu} = \frac{1}{N}\sum_i X_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (X_i - \hat{\mu})^2 = \frac{1}{N}\sum_i X_i^2 - \hat{\mu}^2 \tag{8}$$

The mean value requires thus 2 sums and 1 division per frame, and the standard deviation 3 sums, 1 product and 1 division. Note that this value is independent of the number of frames considered, since they are performed using a sliding window. With all this, for each frame, and considering the four selected features, the number of simple operations required to compute them is equal to $128+2\cdot 3+2\cdot 5=144$.

Although eq. (6) allowed us to estimate the number of simple operations required by a MLP, it is convenient to consider a more conservative estimation, such as:

$$N_{op} = W\,(2L+2M+1)+2M-1+20W \tag{9}$$

Note that this is Eq. (6) with an additional number of operations equal to $20W$. This is necessary to compute the logsig activation function (the logarithm is tabbed in the DSP).

For the speech/non-speech task, where $L=2$, $W=3$ and $M=1$, the number of simple operations needed per frame has been found to be 82. Similarly, for the clean/noisy speech classifier, the number of simple operations required is 55. This takes a total of $144+82+55=281$ operations needed by the whole classification algorithm per frame.

The summarized results suggest that, when properly tailored, the proposed MLP can be feasible implemented on the DSP aiming at classifying into the three classes of interest.

5. DISCUSSION

This paper has proposed the use of a two-layer NN-based sound classifier for a hearing aid to distinguish among speech in quiet, speech in noise and noise. The reason to adopt this two-layer strategy is that the relative importance of each particular probability of classification is different (e.g., the speech/non-speech discrimination is more critical than the clean/noisy speech classification).

With the proposed system, the probability of correct classification obtained for the speech/non-speech task is equal to 86.7%. For the clean/noisy speech task this probability is equal to 88.4%. A brief study of the computational complexity associated to this kind of implementation concludes that, in spite of the doubts related to the feasibility of neural networks for being used in hearing aids, a proper design makes it possible its implementation.

REFERENCES

[1] G. Keidser, "The relationships between listening conditions and alterative amplification schemes for multiple memory hearing aids," *Ear Hear*, vol. 16, pp. 575–586, 1995.

[2] ——, "Selecting different amplification for different listening conditions," *J. of the American Academy of Audiology*, vol. 7, pp. 92–104, 1996.

[3] M. Büchler, "Algorithms for sound classification in hearing instruments," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 2002.

[4] P. Nordqvist and A. Leijon, "An efficient robust sound classification algorithm for hearing aids," *J. Acoustic Soc. Am.*, vol. 115, no. 6, pp. 3033–3041, 2004.

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

[6] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, October 2002.

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

[8] S. Davis and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, August 1980.

[9] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Int. Symp. Music Information Retrieval (ISMIR)*, 2000.

[10] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[11] J. Saunders, "Real time discrimination of broadcast speech/music," in *ICASSSP*, 1996, pp. 993–996.

[12] E. Batlle, H. Neuschmied, P. Uray, and G. Ackerman, "Recognition and analysis of audio for copyright protection: the raa project," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 12, pp. 1084–1091, October 2004.

[13] S. Haykin, *Neural Networks: A comprehensive foundation*. Prentice Hall, 1999.

[14] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, january 2000.

[15] M. Hagan and M. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[16] F. D. Foresee and M. T. Hagan, "Gauss-newton approximation to bayesian learning," *Proceedings of the 1997 International Joint Conference on Neural Networks*, pp. 1930–1935, 1997.

[17] E. Alexandre, L. Cuadra, L. Perez, M. Rosa-Zurera, and F. Lopez-Ferreras, "Automatic sound classification for improving speech intelligibility in hearing aids using a layered structure," in *Lecture Notes in Computer Science*. Springer Verlag, 2006.

[18] N. Wanas, G. Auda, M. S. Kamel, and F. Karray, "On the optimal number of hidden nodes in a neural network," in *IEEE Canadian Conference on Electrical and Computer Engineering*, 1998, pp. 918–921.