16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

SPEECH/MUSIC/NOISE CLASSIFICATION IN HEARING AIDS USING A TWO-LAYER CLASSIFICATION SYSTEM WITH MSE LINEAR DISCRIMINANTS

Enrique Alexandre, Roberto Gil-Pita, Lucas Cuadra, Lorena Álvarez, Manuel Rosa-Zurera

Department of Signal Theory and Communicaciones, University of Alcala
Escuela Politecnica Superior, 28805, Alcala de Henares (Madrid), Spain
phone: + (34) 91 885 6727, fax: + (34) 91 885 6699, 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 implemented approach consists in dividing the classifying algorithm into two layers that make use of two-class 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 audio is noise or music. The complete system results in having three classes, labeled "speech", "noise" and "music". The classification process is carried out by using a *mean squared error linear discriminant*, which provides very good results along with a low computational complexity. This is a crucial issue because hearing aids have *to work at very low clock frequency*. The paper explores the feasibility of this approach thanks to a number of experiments that prove the advantages of using the proposed *two-layer system* rather than a three-classes, *single-layer* classifier.

1. INTRODUCTION

A particular application that would deem as very appreciated by hearing aid users, specially by those eldest, is that in which the hearing aid *itself* classifies the acoustic environment that surrounds him/her, and automatically selects the amplification "program" that is best adapted to such environment ("self-adaptation"). The "manual" approach, in which the user has to identify the acoustic surroundings and choose the adequate program, is very uncomfortable and frequently exceeds the abilities of many hearing aid users. Only about 25% of people owning a hearing aid (a scarce 20% of those that could benefit from hearing aids) wear it because of the unpleasant whistles and/or other amplified noises caused by the surrounding background noise they encounter in their everyday life, and in particular, when moving from one acoustic ambient (for example, speech in quiet) to another different one (say, for instance, a crowded cafe) for which the current program is not fitted (the user thus hears a sudden, uncomfortable amplified noise). This illustrates the necessity for hearing aids able to automatically classify the acoustic environment the user is in. This type of hearing aid could help the user to *improve speech intelligibility*, increasing his/her comfort level and allowing the user to lead a normal life. With respect to comfort, a study [4] with hearing impaired subjects suggests that the automatic switching is deemed useful by most of them, even if its performance is not completely perfect.

The problem becomes more difficult because designing such classifier embedded in a hearing aid is constrained to a very strong limitation: the digital signal processing (DSP) the hearing aid is based on has to work at very low clock frequency in order to minimize power consumption and thus maximize battery life.

Regarding the mentioned issues, the purpose of this work is just the development of a two-layer sound classifier, which programmed on a DSP-based hearing aid, assists it to enhance the user's listening skills, without increasing its computational load. In our previous work [1], it has been shown that the use of a two-layer classification system may provide some interesting advantages. This is just the reason that compels us to explore a divide-and-conquer strategy that leads to a classification systems composed of two more specialized classifying layers. The first one discriminates the input sound into either speech or non-speech, this second category being named "noise" in our work. If the discriminated signal has been found to be speech, a second algorithm in the second layer classifies it into either speech in quiet or speech in noise. The particular class of classifying algorithm we have implemented in the present approach is a *mean squared error linear discriminant*, because it provides very good results along with a low computational complexity, as required by the DSP limitations.

In the effort of making the paper to stand by-itself, after summarizing the important limitations (Section 2) the system suffers from, the paper centers on designing the particular implementation of the *mean squared error linear discriminant* in Section 3. The paper is completed with the experimental work (4), and the discussion of the results (5).

2. DESIGN CONSTRAINTS

As mentioned, DSP-based hearing aids have generally very strong constraints in terms of computational capacity and memory. These restrictions arise mainly from the small size of the hearing aid −specially for the smallest in-the-canal (ITC) or completely-in-the-canal (CIC)−, which must additionally contain a small battery for supplying energy to the DSP. Note that, generally, the DSP has to integrate not only the CPU core but also the A/D and D/A converters, the filter-bank, the RAM, ROM and EPROM memories and some input/output ports. The immediate consequence is that the hearing aid has to work at very low clock frequencies (around 2MHz) in order to minimize the power consumption and thus maximize the life of the battery. Additionally, the restrictions become stronger because a considerable part of

[1]This work has been partially funded by the Comunidad de Madrid/Universidad de Alcalá (CCG06-UAH/TIC-0378, CCG07-UAH/TIC-1572) and the Spanish Ministry of Education and Science (TEC2006-13883-C04-04/TCM).

Feature Number	Name
1	Spectral Centroid (SC) [9]
2	Spectral Roll-Off (RO) [9]
3	Voice2White (V2W) [6]
4	Spectral Flux (Flux) [9]
5	Zero Crossing Rate (ZCR) [9]
6	Short-Time Energy (STE) [9]
7	Percentage of Low Energy Frames (LFE) [8]
8	High Zero Crossing Rate Ratio (HZCRR) [7]
9	Low Short-Time Energy Rate (LSTER) [7]
10	Spectral Flatness Measure (SFM) [3]
11	Mel Frequency Cepstral Coefficients (MFCC) [5]
12	Loudness (Ldn)
13	Spectral Crest Factor (SCF)
14	Bandwidth (BW)

Table 1: List of the 14 considered features for the classification process.

the DSP computational capabilities are already being used for running the algorithms aiming to compensate the acoustic losses. For instance, the filter-bank requires about 50% of the DSP computation time. Therefore, the design of any automatic sound classifier is strongly constrained to the use of the remaining resources of the DSP: roughly speaking, the computational power available does not exceed 3 MIPS, with only 32 Kbytes of internal memory. The time/frequency decomposition is performed by using an integrated Weighted Overlap-Add filter-bank, with 64 frequency bands.

To complete this brief section it is worth mentioning that the complete implementation of the hearing aid itself is out of the scope of this paper, whose purpose is, as pointed out in the Introduction, to select a reduced number of signal-describing features to be programmed on a DSP for automatic sound classification.

3. THE PROPOSED SYSTEM

It basically consists of a feature extraction block, and the aforementioned classifier based on a mean squared error linear discriminant.

3.1 Feature Extraction

For being digitally processed, the input audio signal is segmented into frames with a length of 512 samples (23.22 ms for the considered sampling frequency), and with no overlap between adjacent frames. Then, a Discrete Cosine Transform (DCT) is computed [2], and all the considered features are calculated. Finally, the mean and standard deviation values are computed every 2 seconds in order to soften the values.

A total of 14 different, sound-describing features, listed in table 1, has been considered in the present approach. More detailed considerations about them will be done later on.

3.2 Classification System

A Mean Squared Error (MSE) linear discriminant has been chosen, as mentioned in the Introduction, 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 has been computed in the previous stage:

$$g = f\left(b + \sum_{n=1}^{L} x_n w_n\right) \tag{1}$$

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:

$$\mathbf{Q} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{L1} & x_{L2} & x_{L3} & \dots & x_{LN} \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} \tag{2}$$

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 classifier as:

$$\mathbf{V} = \begin{pmatrix} w_{11} & w_{21} & \dots & w_{L1} & b_1 \\ w_{12} & w_{22} & \dots & w_{L2} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{1C} & w_{2C} & \dots & w_{LC} & b_C \end{pmatrix} \tag{3}$$

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

The output of the classifier can be defined as:

$$\mathbf{Y} = \mathbf{V} \cdot \mathbf{Q} \tag{4}$$

Y being a matrix with C rows and N columns.

The error can be defined as:

$$\mathbf{E} = \mathbf{Y} - \mathbf{T} = \mathbf{V} \cdot \mathbf{Q} - \mathbf{T} \tag{5}$$

where T represents a $C \times N$ matrix containing the target classes for each input pattern. If we define the mean squared error (MSE) as:

$$MSE = \frac{1}{NC} \sum_{n=1}^{N} \sum_{c=1}^{C} e_{cn}^2 \tag{6}$$

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

$$\mathbf{V} = \mathbf{T} \cdot \mathbf{Q}^T \cdot \left(\mathbf{Q} \cdot \mathbf{Q}^T\right)^{-1} \tag{7}$$

4. EXPERIMENTAL SETUP

Prior to the description of the experiments carried out and the discussion of the corresponding results it is convenient: 1) to describe the database used, and 2) to define a tool able to measure how accurate the results are.

4.1 Database Used

The sound database we have 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, music and noise. 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 were both vocal and instrumental. 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.

4.2 Relative error analysis

When comparing the results from different experiments one wonders which of the differences observed in the probability of correct classification achieved by the different approaches may be considered as statistically significant or not. In this respect it is important to analyze the relative error $\varepsilon_{\hat{P_{CC}}}$ of the estimator of probability of correct classification, $\hat{P}_{CC}$, with a given confidence interval ξ, which is given by [10]:

$$\varepsilon^2_{\hat{P_{CC}}} \leq \frac{(1-P_{CC}) \cdot \left[Q^{-1}(\xi/2)\right]^2}{M \cdot P_{CC}} \tag{8}$$

where P_{CC} is the probability to be calculated, M is the number of elements in the test set and $Q^{-1}(x)$ the complementary error function defined as

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right) dt \tag{9}$$

For our application, and assuming a confidence interval $\xi = 0.99$, and probabilities around 0.95, the relative error for the estimation is $\varepsilon_{\hat{P_{CC}}} < 10^{-4}$. That is, a 0.01% relative error should be expected. A more conservative value of 0.1% will be considered, meaning that only those differences between probabilities of correct classification above this value will be considered as significative.

5. THE COMPLETE SYSTEM: RESULTS

The batch of experiments have been carried as follows:

- Each possible combination of features has been considered. Since there are 14 available features, the total number of combinations equals $\sum_{k=1}^{14} \binom{14}{k} = 2^{14} - 1$.
- For each combination of features, the classifier has been trained, and the error probability for the validation set has been evaluated.
- For each number of features, the best result in terms of probability of error for the validation set has been chosen. Finally, the probability of error for the *test* set has been computed by using this combination of features.

Aiming at exploring the advantages of the proposed approach we have compared it with the single-layer classifier described below.

5.1 One-layer classifier

This approach consists in using a single three-classes classifier to distinguish among speech, music and noise. Figure 1 shows the results obtained for the classification of sounds among speech, noise and music. It illustrates the error probabilities obtained for both the validation and the test sets for each possible number of selected features, m. Note that the best result for the validation set is obtained for $m = 11$, corresponding to a probability of error with the test set equal to 4.80%.

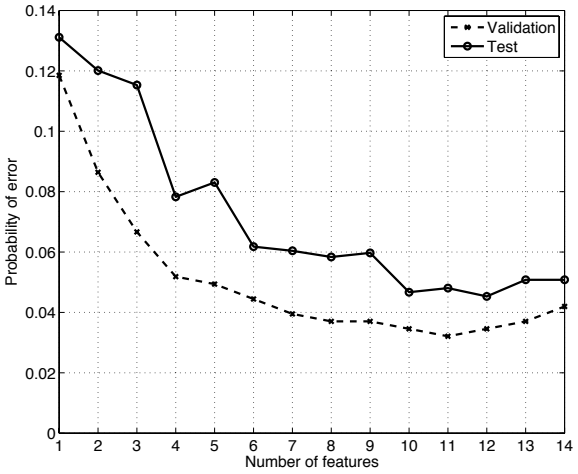

Figure 1: Probabilities of error for the validation (dashed line) and test (solid line) sets achieved for different numbers of input features by the 3-classes classifier.

Number of features	pVal	pTest	Features used
1	11.85	13.11	11
2	8.64	12.01	4, 11
3	6.67	11.53	3, 4 ,11
4	5.19	7.82	4, 6, 11, 14
5	4.94	8.30	4, 6, 11, 12, 14
6	4.44	6.18	4, 6, 10, 11, 13, 14
7	3.95	6.04	4, 5, 6, 10, 11, 13, 14
8	3.70	5.83	4, 5, 6, 10, 11, 12, 13, 14
9	3.70	5.97	3, 4, 5, 6, 9, 10, 11, 13, 14
10	3.46	4.67	2, 3, 4, 5, 6, 10, 11, 12, 13, 14
11	**3.21**	**4.80**	**2, 3, 4, 5, 6, 9 10, 11, 12, 13, 14**
12	3.46	4.53	1, 2, 3, 4, 5, 6, 9 10, 11, 12, 13, 14
13	3.70	5.08	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14
14	4.20	5.08	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Table 2: Probabilities of error for the validation and test sets, and features used for the single-layer speech/music/noise classifier.

These results are shown in Table 2, together with the features selected for each case. Note that the MFCCs (feature 11) is always present, the same for the spectral flux (feature 4). The less-used features are the HZCRR, LEF and SC.

For clarity, Figure 2 illustrates the histogram of the features selected.

5.2 The proposed two-layer classifier

This approach consists in using two cascaded 2-class classifiers. The first classifier aims to distinguish between speech and non-speech, while the second one would classify he latter in either music or noise. This approach has the advantage of making use of two more-specialized classifiers which can easier be trained and implemented.

Figures 3 and 4 show the results obtained for the validation and test sets as a function of the different numbers of

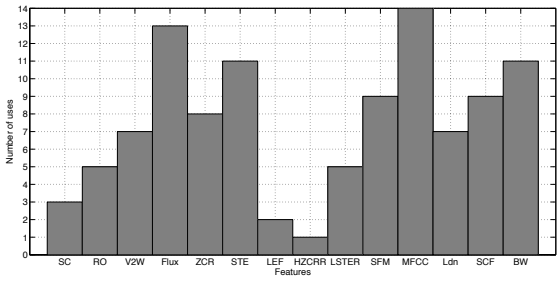

Figure 2: Histogram of the selected features for the single-layer speech/music/noise classifier

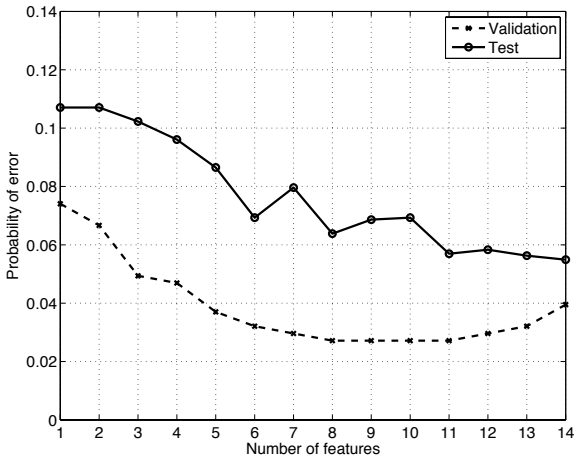

Figure 3: Probabilities of error achieved by the speech/non-speech classifier for the validation (dashed line) and the test (solid line) sets for different numbers of input features.

input features. Note that the second layer classifier (music vs. noise) achieves a zero probability of error for the validation set if the number of features is higher than 2.

For the sake of clarity, the best set of selected features (according to the probability of error for the validation set), for each number of input features, has been listed in Tables 3 and 4.

5.3 1-layer vs. 2-layers

Given the results presented in the previous sections, it is interesting to wonder whether it is worth using a two-layer classifier for the proposed problem. Two factors need to be taken into account: the performance of the classifier and its computational complexity.

Regarding the computational complexity, it is interesting to discuss the difference between using a single 3-class classifier, or two 2-class, cascaded classifiers. Recalling equation 4, the MSE linear discriminant classifier requires $C \cdot (L+1)$ sums and multiplications for each input pattern, with L being the number of input features and C the number of output classes. For the case of a single layer, three-classes classifier, this number turns into $3 \cdot (L+1)$. If only 2 classes are being considered, there is no need for two outputs but only one,

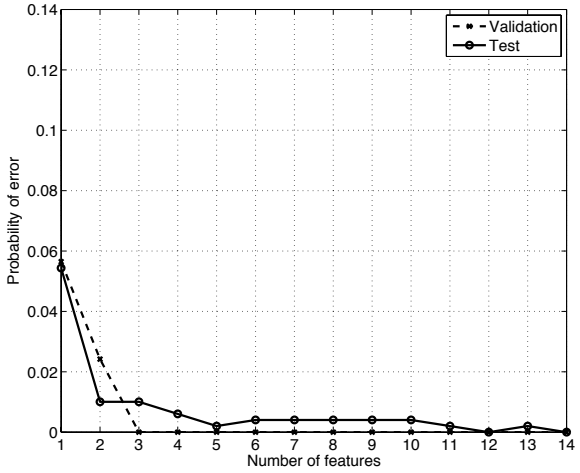

Figure 4: Probabilities of error achieved by the music/noise classifier for the validation (dashed line) and the test (solid line) sets for different numbers of input features.

Number of features	pVal	pTest	Features used
1	7.41	10.71	11
2	6.67	10.71	8, 11
3	4.94	10.23	1, 11 ,12
4	4.69	9.61	2, 11, 12, 13
5	3.70	8.65	2, 5, 10, 11, 13
6	3.21	6.93	1, 4, 6, 10, 11, 14
7	2.96	7.96	2, 5, 7, 10, 11, 13, 14
8	2.96	6.38	1, 2, 4, 6, 8, 10, 11, 14
9	2.96	6.86	1, 2, 4, 6, 7, 9, 10, 11, 14
10	2.96	6.93	1, 2, 4, 6, 7, 8, 9, 10, 11, 14
11	**2.72**	**5.70**	**1, 2, 3, 4, 6, 7 9, 10, 11, 12, 14**
12	2.96	5.83	1, 2, 3, 4, 6, 7, 8 9, 10, 11, 12, 14
13	3.21	5.63	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14
14	3.95	5.49	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Table 3: Probabilities of error for the validation and test sets, and features used for the speech/non-speech classifier.

since it is equivalent to decide based on the maximum value of two outputs or on the thresholded value of only one output. With this in mind, a two-classes classifier would require $L+1$ sums and multiplications, and thus a two-layer classifier would imply only $2 \cdot (L+1)$ sums and multiplications, 50% less than for the three-classes classifier.

As it was commented above, hearing aids suffer from strong computational complexity constraints. The best solution achieved by the 3-classes classifier returns a probability of error equal to 4.80%, with 11 features. This number of features is, however, excessive given their computational cost. A number of features like 2 or 3 seems much more feasible. With this constraint, the probability of error achieved by this classifier equals 11.53% (with 3 features).

On the other hand, for the 2-layers classifier, the best option is to use 11 features for the speech/non-speech task, and 3 features for the music/noise task. This combination returns a probability of error equal to 6.38% and 1.01% for the speech/non-speech and the music/noise problems respectively. Like for the previous case, this number of features is

Number of features	pVal	pTest	Features used
1	5.65	5.43	11
2	2.42	1.01	10, 11
3	**0.00**	**1.01**	**6, 10 ,11**
4	0.00	0.60	10, 11, 12, 14
5	0.00	0.20	10, 11, 12, 13, 14
6	0.00	0.40	9, 10, 11, 12, 13, 14
7	0.00	0.40	8, 9, 10, 11, 12, 13, 14
8	0.00	0.40	7, 8, 9, 10, 11, 12, 13, 14
9	0.00	0.40	6, 7, 8, 9, 10, 11, 12, 13, 14
10	0.00	0.40	5, 6, 7, 8, 9, 10, 11, 12, 13, 14
11	0.00	0.20	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
12	0.00	0.00	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
13	0.00	0.20	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
14	0.00	0.00	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Table 4: Probabilities of error for the validation and test sets, and features used for the music/noise classifier.

excessive, and should be reduced. If we consider only 3 features for each classifier, these probabilities become 10.23% and 1.01%. However, these 3 features are not the same for both classifiers (in fact, only one of them is common), and thus the comparison would not be fair. If the same three features are considered for both layers, the results obtained are 10.23% for the speech/non-speech classification and 1.41% for the music/noise classification. This implies a final probability of error among the three classes equal to 10.65%, slightly lower than the 11.53% achieved with the one-layer, three-classes classifier.

However, as it was commented above, the single layer classifier has a higher computational complexity. For the same computational complexity, it is possible to use $(3L+1)/2$ input features instead of L with the two-layer classifier. This means that while the three-classes classifier would use only 3 input features, the two-layers classifier could use 5 input features while maintaining the same computational complexity for the classifier. With this, the probabilities of error achieved with the two-layers classifier would be 8.65% for the speech/non-speech problem and 0.60% for the music/noise classifier (considering that the same features are used for both classifiers). This returns an overall mean probability of error equal to 8.83%.

6. DISCUSSION

In this paper we have explored a two-layer, MSE-linear-discriminant-based classifier to be implemented on 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.

The kind of classifying algorithm explored here has been the MSE linear discriminant that exhibits very good results and assists in the goal of using efficiently the scarce computational resources. The particular structure we have adopted is based on a divide-and-conquer strategy that leads to a layered structure with 2 layers, 2 binary classifiers, and three classes (speech, music and noise). The first layer centers on classifying the input signal into either speech or non-speech. The second layer discriminates audio files between music and noise.

In order to check the results, we have carried out a number of experiments to compare the proposed 2-layer, 3-class, MSE-linear-discrimant-based approach with that corresponding to a three-classes, single-layer classifier. This comparison has been evaluated in terms of both computational complexity and performance. The experiments prove that the two-layer approach presents a lower computational complexity in terms of number of sums and multiplications required to obtain an output. For a similar computational complexity, the single-layer system obtains an error probability equal 11.53%, while the dual-layer system reduces the error probability down to 8.83%.

REFERENCES

[1] E. Alexandre, L. Cuadra, L. Álvarez, M. Rosa-Zurera, and F. Lopez-Ferreras. Two-layer automatic sound classification system for conversation enhancement in hearing aids. *Integrated Computer-Aided Engineering*, 15(1):85–94, 2008.

[2] Enrique Alexandre, Manuel Rosa, Lucas Cuadra, and Roberto Gil-Pita. Application of fisher linear discriminant analysis to speech/music classification. In *AES 120th Convention, Preprint #6678*, 2006.

[3] Eloi Batlle, Helmut Neuschmied, Peter Uray, and Gerd Ackerman. Recognition and analysis of audio for copyright protection: the raa project. *Journal of the American Society for Information Science and Technology*, 55(12):1084–1091, October 2004.

[4] M.C. Büchler. *Algorithms for sound classification in hearing instruments*. PhD thesis, Swiss Federal Institute of Technology, Zurich, 2002.

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

[6] Enric Guaus and Eloi Batlle. A non-linear rhythm-based style classification for broadcast speech-music discrimination. In *AES 116th Convention*, 2004.

[7] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on speech and audio processing*, 10(7):504–516, October 2002.

[8] J. Saunders. Real time discrimination of broadcast speech/music. In *ICASSSP*, pages 993–996, 1996.

[9] Eric Scheirer and Malcom Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *ICASSP*, 1997.

[10] K.S. Shanmugam and P. Balaban. A modified monte carlo simulation technique for the evaluation of error rate in digital communication systems. *IEEE Transactions on Communications*, COM-28(11):1916–1924, November 1980.