A NEW RELATIVISTIC VISION IN SPEAKER DISCRIMINATION

S. Ouamour*, M. Guerti#, H. Sayoud*

**USTHB University, Institut d'Electronique.*
# Ecole Nationale Polytechnique.
USTHB, Institut d'Electronique, BP 32 Bab-Ezzouar, Alger, Algeria.
* *siham.ouamour@gmail.com* * *halim.sayoud@gmail.com*

ABSTRACT

The present paper deals with the task of speaker discrimination using a new relativistic approach. Speaker discrimination has two practical applications: speaker verification and audio document indexing. In such applications, the speaker model is extracted directly from speaker's own speech signal as well as using speaker's own features. However, such a model can be rigid, inaccurate and not appropriate in fluctuating environments where a change in the recording conditions may occur. For instance, during telephone talks, the vocal features for the same speaker may change considerably. And hence, a new relative speaker model is introduced. The new model is based on a relative characterization of the speaker, called Relative Speaker Characteristic (RSC). RSC consists in modeling one speaker relative to another, meaning that each speaker model needs both its speech signal and its competing speech (speech of the speaker to be compared with). This investigation shows that the relative model, used as input at a neural network classifier, optimizes the training of the classifier, speeds up its learning time and also enhances the discrimination accuracy. The experiments of speaker discrimination are done on two different databases: Hub4 Broadcast-News database and a telephonic speech database by using a Multi-Layer Perceptron (MLP) with several input characteristics. Results indicate that the best characteristic is the RSC, when compared to other reduced features evaluated in the same manner.

RÉSUMÉ

Le présent papier s'intéresse à la tâche de discrimination du locuteur en utilisant une nouvelle approche relativiste. La discrimination du locuteur a deux applications pratiques : la vérification du locuteur et l'indexation des documents audio. Dans de telles applications, le modèle du locuteur est extrait directement de son propre signal de parole et en utilisant ses propres caractéristiques. Mais ce type de modèle peut être rigide, imprécis et non approprié dans les environnements fluctuants, où un changement dans les conditions d'enregistrement risque d'arriver. Par exemple, durant les communications téléphoniques, les caractéristiques vocales pour un même locuteur peuvent changer considérablement. Ceci nous a incité à introduire une nouvelle modélisation relative du locuteur. Ce nouveau modèle est basé sur une caractérisation relative du locuteur, appelée Caractéristique Relative du Locuteur (RSC). La RSC consiste à modéliser un locuteur relativement à un autre ; ce qui signifie que pour chaque modèle de locuteur nous avons besoin en même temps de son signal de parole et de son signal dual (signal de parole du locuteur à faire comparer avec). Cette étude montre que le modèle relatif, utilisé comme entrée d'un classifieur connexionniste, permet d'optimiser l'entraînement du classifieur, d'accélérer son temps d'apprentissage et d'améliorer aussi la précision de discrimination. Les expériences de discrimination de locuteur sont effectuées sur deux bases de données : Hub4 Broadcast- News et une base de données d'enregistrements téléphoniques, en employant un Perceptron Multi-couches (MLP) avec plusieurs caractéristiques d'entrée. Les résultats indiquent que la meilleure caractéristique est la RSC, comparativement à d'autres caractéristiques réduites qui sont évaluées de la même manière.

1. INTRODUCTION

Speaker discrimination consists in checking whether two different pronunciations (speech signals) are uttered by the same speaker or by two different speakers (Rose, 2007). This research domain has several applications such as automatic speaker verification, speech segmentation (Meignier, 2006) (Meignier, 2002) or speaker based clustering. All these tasks can be performed either by generative classifiers or by discriminative classifiers, but in practice the second type is simpler and more reliable for short training cases: it consists in a simple comparison between the speech segments.

One method of comparing the speech utterances is to extract the vocal characteristics from each speaker signal, in order to detect the degree of similarity between them.

While fingerprints and retinal scans are more reliable means of authentication, speech can be seen as a non-evasive biometric key that can be collected with or without the person's knowledge or even transmitted over long distances via telephone. Furthermore, a person's voice cannot be stolen, forgotten or lost. Thus, speaker discrimination allows for a secure and efficient method of authenticating speakers. However, existing approaches are not robust enough in noisy environment or for telephonic speech. Any new model must therefore improve the reliability of existing discriminative systems, without altering their architectures.

To address the above issue, a new relativistic characteristic is proposed. The reliability of the new approach is also compared to several other reduced features and thereby show its performance. Experiments show that the use of the new characteristic at the input of a discriminative classifier enhances the discrimination quality. The new approach is called "Relative Speaker Characteristic (RSC)." Basically, the introduction of the relative notion in speaker modelization allows getting a flexible relative speaker template, more suitable for the task of speaker discrimination in difficult environments.

The format of this paper is as follows: In section 2, we give the motivation of this research work and describe some related works. Section 3 introduces the Relativity in speaker discrimination. Section 4 describes the RSC based Neural Network (NN) used for the task of speaker discrimination. Experiments of Speaker Discrimination are presented in section 5 and finally a short conclusion is given.

2. MOTIVATION AND RELATED WORKS

2.1 Speaker recognition, applications and some problems

Speaker recognition is the ability to recognize the speaker, by using the vocal characteristics of his or her speech signal. Speaker recognition is divided into several specialties: speaker identification, speaker verification, speaker indexing and speaker discrimination.

- Speaker identification is the ability to identify the identity of a speaker among others;
- Speaker verification is the process of accepting or rejecting the identity claim of a speaker;
- Speaker indexing consists in segmenting and labeling a multi speaker audio document into homogenous segments containing only one speaker;
- Speaker discrimination is the ability to recognize whether two utterances come from the same or different speakers. This field is an important component of segmenting an audio stream into meaningful subunits, because the location of the speaker changes is crucial for dialogue understanding. Speaker discrimination is also related to speaker verification, but this last process is based on prior knowledge about a limited number of speaker identities, whereas in speaker discrimination, only knowledge about the speech signal is provided.

Speaker recognition has several practical applications in voice dialling, banking transactions by telephone, database access services, voice mail, biometric secure access, and forensic applications.

The problems encountered in speaker recognition are usually due to the intra-speaker variability of the speech, effect of noise and reduction of the spectral bandwidth in telephonic speech: [300-3400Hz]. These problems led to the choice of two types of speech databases during for the experiments, namely Hub4 Broadcast-News for the corrupted speech and telephonic calls for the reduced spectral bandwidth, in order to evaluate the proposed approach.

2.2 Some feature extraction and reduction techniques

Different techniques were developed for the task of features reduction during the last few years. In 1974, Attal (Atal, 1974) used low dimension Auto Regressive coefficients. In 1992, Bennani (Bennani, 1992) investigated the use of mean and eigenvectors of the covariance matrix. Then in 1995, Reynolds proposed the use of the covariance diagonal (Reynolds, 1995) for modeling the Gaussian Mixture Models (GMMs) and in 1995 Bonastre used the sub-bands combination (Bonastre, 1997) in order to select the best spectral bands. Later on, in 2000, Magrin-Chagnolleau conducted an investigation on alternative speech features using Line Spectrum Pairs (LSP), Time- Frequency Principal Components (TFPC) and Discriminant Components of the Spectrum (DCS) for the task of speaker characterization, but his experiments did not succeed in evidencing a benefit of alternate features over classical cepstral coefficients (Magrin, 2000).

Even in the field of speech recognition, Wang indicated in 2003 that although Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are the two most popular independent feature extraction methods, the drawback of independent feature extraction algorithms is that their optimization criteria are different from the classifier's minimum classification error criterion, which may cause inconsistency between feature extraction and the classification stages of a pattern recognizer (Wang, 2003).

Recent works in speaker recognition have demonstrated the advantage of modeling stylistic features in addition to traditional cepstral features (Ferrer, 2006), but the extraction of such features remains difficult in practice.

In 2006, Mami introduced the speaker representation by location in a reference space (Mami, 2006), which is a new technique of speaker recognition and adaptation.

After a thorough investigation on the optimal spectral resolution for speaker characterization Sayoud showed that the spectral parameterization of 37 Mel Frequency Spectral

Coefficients (MFSC) was optimal, implying that high spectral resolutions are interesting in speaker discrimination. (Sayoud, 2000; Sayoud, 2006) However the high dimensionality of the corresponding covariance makes the training step considerably difficult when short speech segments (few training data) were used for the segmentation task. One way to overcome this dimensional issue is to use a reduced and relative characteristic. For this reason, a new relative characteristic called RSC derived from the MFSC coefficients would be used for the task of speaker discrimination.

This relativity approach reduces the features dimension, optimizes the neural network training and tries to improve the speaker discrimination accuracy, without modifying the classifier architecture or without changing the input features.

In fact, the principle is to exploit the usual features and compute the covariance matrix for the whole utterance. We redo the same process for the second utterance to compare. After that, we compute the RSC characteristic (as we will see in section 3), and extract the diagonal vector which will replace the old features at the input of the classifier.

3. RELATIVITY AND DISCRIMINATION

3.1 Introduction

In this research work, we try to introduce a new approach of speaker recognition based on relativist discrimination. This new approach leads to a new way of classification, which can be used in some applications as speaker discrimination, speech recognition, speech segmentation and so on. Instead of drawing the boundaries between the different classes (figure 1-a), the relativity based method consists in analyzing all the possible combinations between all couples of examples and then, keeping only the minimal-distance combinations, which should indicate the examples having a similarity with the corresponding relative reference. All the examples linked to a relative reference are considered having the same type (fig. 1-b).

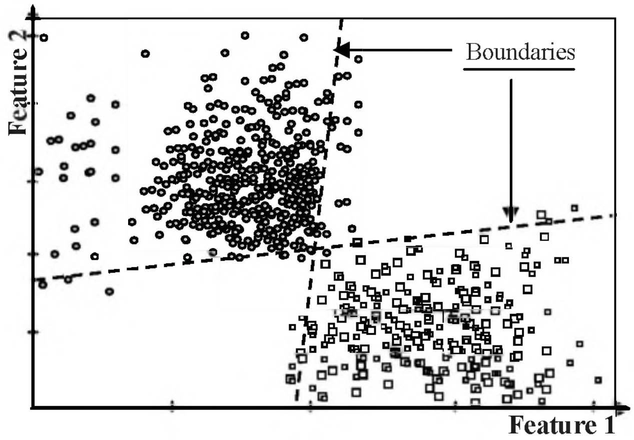


Figure 1-a: Absolute Linear classification:
Absolute boundaries are set between the two classes of examples. Features 1 and 2 represent two pertinent features of the examples.

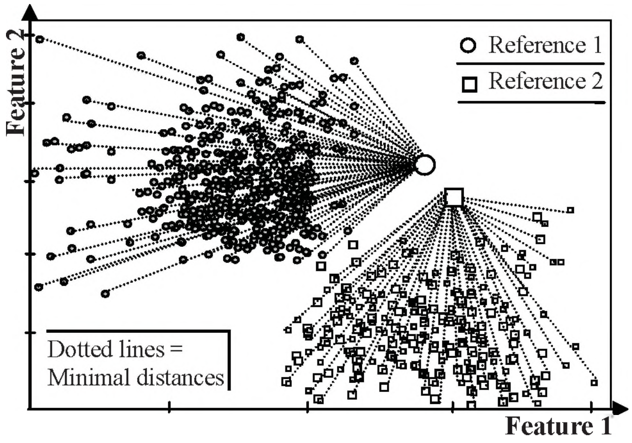


Figure 1-b: Relative Classification:
Relative links are set between the examples and the two references. A relative discrimination is made with respect to the references. No boundaries are set but the examples are relatively classified according to their minimal distances from the two references.

3.2 Some statistical similarity measures used in Speaker discrimination

A classical discrimination method based on mono-Gaussian models uses some measures of similarity, which are called Second Order Statistical Measures. These measures are used in order to recognize the speaker at each segment of the speech signal.

We recall below the most important properties of this approach (Gish, 1990; Bimbot, 1995; Bonastre, 1997). Let $\{x_t\}_{1\leq t\leq M}$ be a sequence of M vectors resulting from the *P*-dimensional acoustic analysis of a speech signal uttered by speaker ***x***. These vectors are summarized by the mean vector $\bar{x}$ and the covariance matrix X:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \qquad (1)$$

and $$X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \qquad (2)$$

Similarly, for a speech signal uttered by speaker ***y***, a sequence of N vectors $\{y_t\}_{1\leq t\leq N}$ can be extracted. By assuming that all acoustic vectors extracted from the speech signal uttered by speaker ***x*** are distributed like a Gaussian function, the likelihood of a single vector y_t uttered by speaker ***y*** is

$$G(y_t/\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{-(1/2)(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \qquad (3)$$

If all vectors y_t are assumed to be independent observations, the average log-likelihood of $\{y_t\}_{1\le t\le N}$ can be written as

$$\overline{L}x(y_1^N) = \frac{1}{N}\log G(y_1...y_N|\boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t|\boldsymbol{x}) \quad (4)$$

We also define the minus-log-likelihood $\psi(\boldsymbol{x}, y_t)$ which is equivalent to similarity measure between vector y_t (uttered by $\boldsymbol{y}$) and the model of speaker $\boldsymbol{x}$, so that

$$Arg\min_x \psi(\boldsymbol{x}, y_t) = Arg\max_x \; G(y_t / \boldsymbol{x}) \quad (5)$$

And hence,

$$\psi(\boldsymbol{x}, y_t) = -\log \; G(y_t / \boldsymbol{x}) \quad (6)$$

The similarity measure between test utterance $\{y_t\}_{1\le t\le N}$ of speaker $\boldsymbol{y}$ and the model of speaker $\boldsymbol{x}$ is then

$$\psi(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x}, y_1^N) = \tfrac{1}{N}\sum_{t=1}^{N}\psi(\boldsymbol{x}, y_t) \quad (7)$$

$$= -\overline{L}x(y_1^N) \quad (8)$$

After simplifications (Sayoud, 2003b, Bimbot, 1995), we obtain

$$\psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1}) + (\overline{y}-\overline{x})^T X^{-1}(\overline{y}-\overline{x})\right] - 1 \quad (9)$$

This measure is equivalent to the standard Gaussian likelihood measure defined in (Bimbot, 1995; Sayoud, 2003). A variant of this measure called μ_{Gc} is deduced from the previous one by neglecting the third term:

$$\mu_{Gc}(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{P}(\overline{y}-\overline{x})^T X^{-1}(\overline{y}-\overline{x}) \quad (10)$$

3.3 Notion of RSC (Relative Speaker Characteristic)

Natural techniques of discrimination, as those used by human beings, are based on relative assessments or comparisons of something/ somebody with respect to a referential object or person in one's memory. For concreteness, everyone can easily make a discrimination between himself and another person, only by observing his relative height (relative to a model in memory) and deduce if the person near him is an adult or a child (figure 2).

The relative statistics, between the utterances of 2 speakers, represent the statistical features of one speaker relatively to another one considered as a reference speaker. The previous formula 9 gives a similarity measure between a speech signal uttered by a speaker $\boldsymbol{y}$ and the reference model of the speaker $\boldsymbol{x}$:

$$\psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1}) + (\overline{y}-\overline{x})^T X^{-1}(\overline{y}-\overline{x})\right] -$$

Due to the fact that the between-variability of the mean vector is low, and is insignificant in noisy or telephonic environment (Sayoud, 2000) we can write:

$\overline{y} \approx \overline{x}$ (i.e. the variability of the mean is negligible).

Moreover, if x and y represent the same speaker, then this approximation is justified. In the other hand, even if the speakers are different, we can make them equal by a special normalization (e.g. normalization by the mean).

So, according to this hypothesis, the approximated similarity measure becomes:

$$\psi^*(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1})\right] - 1 \quad (11)$$

$\frac{\det(Y)}{\det(X)} = \det(Y / X)$, where Y / X represents the expression $Y.X^{-1}$, and hence:

$$\psi^*(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{P}\left[-\log(\det(Y / X) + tr(Y / X \;))\right] - 1 \quad (12)$$

And if we denote the ratio Y/X by $\Re(\boldsymbol{x,y})$ or simply $\Re$, then

$$\psi^*(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{P}\left[-\log(\det(\Re) + tr(\Re))\right] - 1 \quad (13)$$

The $\Re$ ratio is **Relative Speaker Characteristic** (RSC)

$$RSC(\boldsymbol{x}, \boldsymbol{y}) = \Re = \frac{Y}{X} = Y * X^{-1} \quad (14)$$

Hence, $\psi^*(\boldsymbol{x}, \boldsymbol{y})$ appears to be a function of the RSC.

3.4 Importance of the diagonal

Let us define a modified similarity measure $\psi^{\#}$ as follows:

$$\psi^{\#}(\boldsymbol{x}, \boldsymbol{y}) = P.(\psi^*(\boldsymbol{x}, \boldsymbol{y}) + 1) \quad (15)$$

After simplification,

$$\psi^{\#}(\boldsymbol{x}, \boldsymbol{y}) = \left[-\log(\det(\Re) + tr(\Re)\right] \quad (16)$$

The two similarity measures $\psi^{\#}$ and ψ^* are proportional and physically equivalent. We will see now the principal components of this modified measure (formula 16). Globally, the value of this measure is closely dependent on the diagonal elements of the $\Re$ matrix. But this dependence is debatable and we can consider four cases.

Case 1: if the two utterances are the same, then the $\Re$ matrix is reduced to the Identity matrix, which confirms the previous statement;

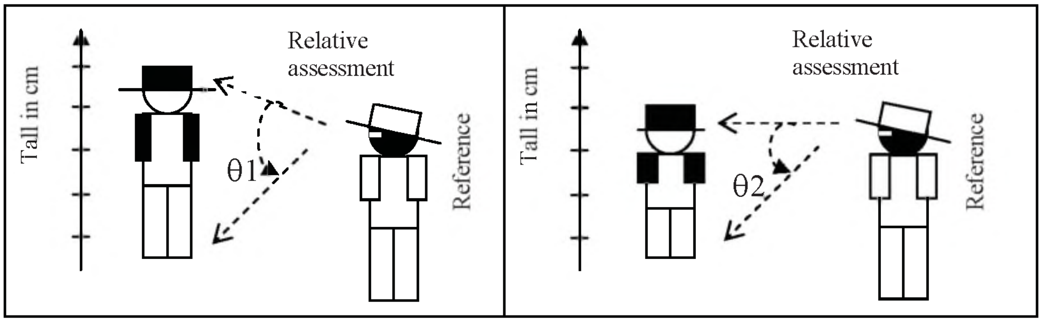


Fig. 2: Relative assessment used in natural human recognition to discriminate between adults and children: in the left side the reference can recognize that the person next to him is an adult; in the right side the reference can recognize that the person next to him is a child. This assessment is made relatively, without using any ruler to measure the person tall.
- θ1 or θ2 is the relative angular tall perceived by the eye -

Case 2: if the two utterances belong to the same speaker, then the $\Re$ matrix is more or less close to the identity matrix even if the non-diagonal elements are non zero. This confirms the previous statement too;

Case 3: if the two utterances belong to different speakers, then the $\Re$ matrix looses the identity form, but if the Speakers' features are not too different, one should retrieve large values on the $\Re$ matrix diagonal (relatively very greater than the non-diagonal elements). The reason is that the two audio signals do have a lot of common acoustic and physiologic characteristics anyway, which are typical to the speech nature of the acoustic signal;

Case 4: if the two audio signals have different types of sources (e.g. one signal is speech and the other is noise or music), then they result in random values in the $\Re$ matrix. The non-diagonal elements of $\Re$ could not be neglected.

Therefore, for the three first cases, more information on the diagonal of the RSC ($\Re$ matrix) is thus obtained. More the similarity between the two signals, the diagonal will be dominant and rich in information. Moreover, if the speech signal is strongly noised or if the two transmission channels are very different we may meet the same problem even if the two speakers are the same.

3.5 RSC pertinence and Symmetry

Let us denote by $\lambda_i|_{i=1..p}$ the eigenvalues of $\Re$ and:
since $\det(\Re) = \prod_i \lambda_i$

and $\mathrm{tr}(\Re) = \sum_i \lambda_i$

we can write :

$$\psi^{\#}(\boldsymbol{x},\boldsymbol{y}) = [-\log(\prod_i \lambda_i) + \sum_i \lambda_i] \qquad (17)$$

$$\text{or } \psi^{\#}(\boldsymbol{x},\boldsymbol{y}) = \sum_i [\lambda_i - \log(\lambda_i)] \qquad (18)$$

if we denote by $\psi_i^{\#}(\boldsymbol{x},\boldsymbol{y})$ the expression $[\lambda_i - \log(\lambda_i)]$ representing the measure part related to the eigenvalues λ_i.

$$\psi_i^{\#}(\boldsymbol{x},\boldsymbol{y}) = [\lambda_i - \log(\lambda_i)] \qquad (19)$$

Then we can write $\psi^{\#}(\boldsymbol{x},\boldsymbol{y}) = \sum_i \psi_i^{\#}(\boldsymbol{x},\boldsymbol{y})$ (20)

The variation of the function $\psi_i^{\#}$ versus λ_i is represented on figure 3. According to figure 3, we can distinguish 2 areas: for $\lambda_i < 1$ (left side) and for $\lambda_i > 1$ (right side). Since the information is focused on the great values of λ_i and since the right side of the figure is more or less linear, it is more accurate to favor the use of eigenvalues greater than 1 resulting in three cases.

Ccase 1: If $\lambda_i > 1$ $\forall$ i,
then we are in the right side of the figure, and the measure is accurate.

Case 2: If the λ_i are > 1 $\forall$ i < ρ,
then we can consider that the dominant information is in the dominant eigenvectors (i < ρ), which leads (with i<ρ) to the same results as for the first case, provided that most of the eigenvalues are superior to 1.

Case 3: If the λ_i are < 1 $\forall$ i,
herein, we are in the left side, and the measure is not linear: varies abruptly with the eigenvalues. This may cause some problems of false rejection.

A new way to unify all these cases, is to consider the two RSC forms: $\Re$ (*x*,*y*) and $\Re$ (*y*,*x*), and integrate them respectively into a new matrix (matrix of matrices).

$$\mathrm{RSC}_{\mathrm{Hybrid}} = [\,[\Re(x,y)]\,,\,[\Re(y,x)]\,] \qquad (23)$$

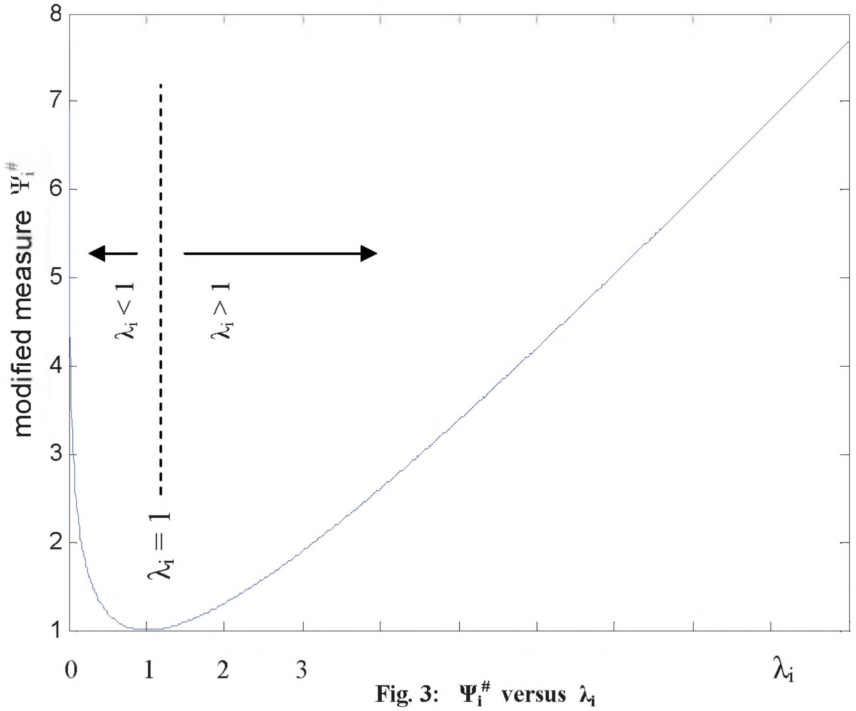


Fig. 3: $\Psi_i^{\#}$ versus λ_i

And since we have shown that the most important information is usually located in the diagonal of the RSC, we propose to use the following relative characteristic, called *symmetric DRSC* ("D" stands for Diagonal):

$$\mathrm{DRSC}\ (\boldsymbol{x},,\boldsymbol{y}) = [\ \mathrm{diag}(\Re\ (\boldsymbol{x},,\boldsymbol{y})) \cup \mathrm{diag}(\Re\ (\boldsymbol{y},\boldsymbol{x}))\ ] \qquad (24)$$

where $\cup$ denotes the concatenation operator.

The DRSC (Diagonal of the RSC) contains enough information normally able to make a correct discrimination between the speakers ***x*** and ***y***.

Its great interest comes from the low dimension of the DRSC vector which allows minimizing the features size and the processing time in particular when using neural networks. For instance if we use acoustic features of 24 coefficients and their derivatives, we should need 2(48x48) = 4608 components in the covariance matrix to exploit, whereas the DRSC input needs only 2(48) = 96 components and which represents only 96/4608 = 2% of the memory space required for the first case. So the simplification, in term of processing time and training data, will be appreciable.

4. USING THE RSC CHARACTERISTIC IN SPEAKER DISCRIMINATION

Knowing the high discriminative capacities of the NNs (neural networks) (Bennani, 1992; Bennani, 1995), we opted for the use of a Multi-Layer Perceptron using the RSC characteristic as input. Experiments of discrimination are done on audio signals, with a speech duration of four seconds in the first experiment and ten seconds, respectively, in the second experiment.

We use the DRSC characteristic as reduced input vector for the NN, which allows us to improve its performance considerably. Furthermore, by using the Relative Speaker Characteristic, we reduce the size of the NN input and the time of training too.

In fact, the NN must have a number of receptive cells equal to the dimension of the example vector (Sayoud, 2003b). Thus, in case of using an input matrix with *PxQ* coefficients (Lee, 1995; Sayoud, 2003b), the number of input receptive cells is equal to *2PQ*.

An example is shown for concreteness:

- In the case of using acoustic features of P coefficients with RSC reduction, the number of input receptive cells is equal to P if we use non-symmetric DRSC, and it is 2P if we use symmetric DRSC.
- But, in the case of using acoustic features of P coefficients, the resulting covariance matrix will have a size of PxP and then P^2 components are required by the classifier.

So, although P^2 components are needed to exploit the classic parameterization, with RSC parameterization only P (or 2P if symmetric) components are required. Such a strong size reduction is interesting since it simplifies the NN architecture, diminishes the required training data set and reduces the learning time.

Concerning the NN architecture, we used Multi layer Perceptrons with 1 or 2 hidden layers and one output neuron. The training is performed by the back-propagation algorithm. The NN output will give then an indication on the correlation between the two utterances. If $NN_{OUTPUT} = 0$ then the two utterances come from the same speaker. If $NN_{OUTPUT} = 1$ then the two utterances belong to different speakers. Concerning the acoustical-spectral analysis of the signal, a segmentation by windows of 35 ms (ensuring the stationarity of the signal) is used in each segment where a spectral analysis is made, in giving a series of MFSC vectors for each segment (Lee, 1995; Sayoud, 2003a).

This vector set goes through a statistical process, which allows extracting the DRSC components in each couple of segments to compare. The DRSC is directly injected to the NN input which will decide whether the two segments belong to the same speaker or not: see figure 4.

5. EXPERIMENTS

5.1 Database and experimental protocol

The aim of our experiments is to check the reliability of the new relative characteristic in speaker discrimination. One part of the experiments concerns the comparison between the DRSC and other existing features such as diagonal of the covariance, mean vector and the first two eigenvectors of the covariance. The other part deals with the investigation of a neural classifier using this new speaker characterization in order to assess its discriminative performance compared to a classical statistical classifier. At the end, a fusion attempt between those classifiers is proposed to further enhance the discrimination accuracy.

Experiments of speaker discrimination are conducted on four databases, as described below:

- Two sub-sets (DB1 and DB2) of “Hub4 Broadcast-News 96” database, containing some recordings from the “CNN early edition” and composed of clean speech, music, telephonic calls, noises, etc. The sampling frequency is 16 kHz. The speech signals are extracted and arranged into segments of about 4 seconds each.
- Two other sub-sets (TB1 and TB2) containing some real telephonic recordings with a sampling frequency of 8 kHz. The duration of each speech segment is about 10 seconds.

In all the databases, the testing examples are different from the training ones.

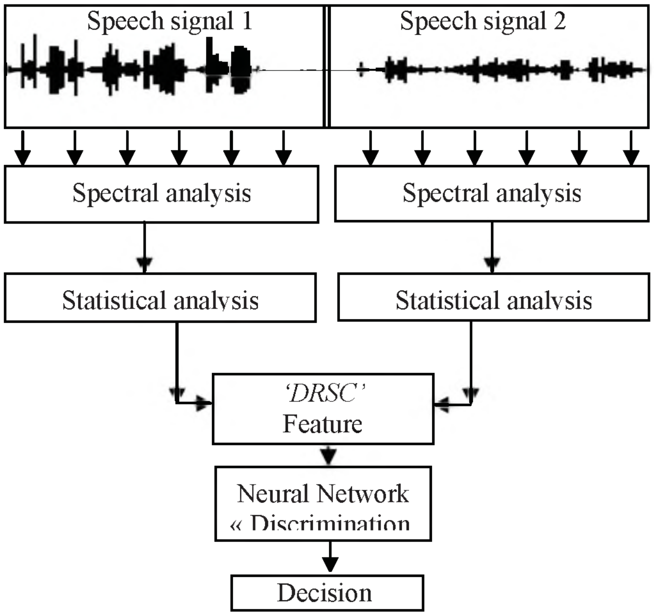


Fig. 4: Relative discrimination between two speech signals.

In addition, the experiments are done with two speech types: The first type is extracted from Hub4 Broadcast-News, which has a large bandwidth of [0-8000Hz] with sometimes some sequences of advertisement, music or noise. The second type is collected from real telephonic conversations, which has a reduced bandwidth of [300-3400Hz]. Usually (not always) speaker recognition is more difficult in telephony, due to the limited bandwidth. However, the presence of noise and music in Hub4 Broadcast-News make the discrimination task rather difficult in this case.

The databases are organized into speaker combinations, namely: pairs of two speech segments to discriminate. The sizes of the different databases are indicated below:

-DB1 contains 14 different speakers (most of them journalists, speaking about the news) organized into 259 speaker combinations for the training and 195 speaker combinations for the test.

-DB2 contains 14 different speakers (most of them journalists, speaking about the news) organized into 518 combinations for the training and 390 combinations for the test.

-TB1 contains 24 different speakers: 12 males and 12 females (speaking by telephone about different topics), organized into 670 speaker combinations for the training and 334 speaker combinations for the test.

-TB2 contains 24 different speakers: 12 males and 12 females (speaking by telephone about different topics), organized into 1340 combinations for the training and 668 combinations for the test.

5.2 Performance Comparison between the RSC and other reduced features

In order to evaluate the different speaker characterizations during the different comparative experiments, we use some common error rates for the performance evaluation. Their definitions are given here below:

- False Alarms (FA): represents the errors in case the system decides that the two speech signals (to compare) do not belong to the same speaker, whereas they really come from the same speaker.

- Missed Detections (MD): represents the errors in case the system cannot detect the difference between two speech signals belonging to two different speakers.

- Equal Error Rate (EER) represents the error of speaker discrimination when the FA ratio is equal to the MD ratio. Then the EER is equal to both FA and MD.

Results of experiments are given in figures 5 and 6, and tables 1 and 2.

Table 1 exposes the different Equal Error Rates (EERs) with their corresponding number of iterations required for the NN training.

These EERs are obtained on a sub-set of Hub4 Broadcast-News database: DB1 (section 5.1), with several speaker characterizations, namely: Diagonal of the Relative Speaker Characteristic (DRSC), diagonal of the covariance, mean vector and the first 2 eigenvectors of the covariance. Results show that the NN using the DRSC characteristic as input gives the best performance with an EER of only 7.20% and the lowest number of iterations for the training (between 1000 and 1500), while by using the diagonal of the covariance as input, the NN causes an EER of 13.90% (the double of that obtained by the DRSC). With the mean vector, the EER is 25.19% and with the first 2 eigenvectors of the covariance the EER is 33.67%. This last one represents the worst discrimination score.

The above experiments are repeated with telephonic database (TB1), with a duration of 10 seconds for each speech signal. The results are summarized in Table 1 below.

Table 1: Equal Error Rates obtained, with different features, on DB1.

FEATURE	Approximate number of iterations during the training	EER %
DRSC	between 1000 and 1500	7.20
Diagonal of the covariance	between 3500 and 4000	13.90
Mean vector	between 6500 and 7000	25.19
The first 2 eigenvectors of the covariance	between 2500 and 3000	33.67

Once again, results confirm the good performance of NNs using the RSC characteristic as input, comparatively to the other characteristics tested on the same conditions. This new relative characteristic associated to a 2-hidden layers NN with 10000 iterations gives an EER of 4.65%, while the other characteristics, tested in the same conditions need a much greater number of iterations for the training, as it is in the case of the mean vector: 200000 iterations (20 times of what is required by the DRSC), and for which the EER is 7.01%.

Concerning the diagonal of the covariance, the EER is 10.94%. And for the eigenvectors, we remark that they do not perform well: their EER is 17.5%.

Table 2: Performances obtained with different features, on TB1

FEATURE	Number of iterations during the training	Learning Rate	EER %
DRSC	10000	0.01	4.65
Diagonal of the covariance	20000	0.005	10.94
Mean vector	200000	0.001	7.01
The first 2 eigenvectors of the covariance	100000	0.001	17.5

In order to give a better presentation of the discrimination results provid 5 and 6, respectively, display the different Receiver-Operating-Characteristic (ROC) curves of the errors for the different types of features evaluated on DB1 and TB1. It is seen that the NN using the RSC characteristic has got the best performance since it has considerably reduced the EER, followed by the diagonal of the covariance or the mean vector and finally by the first 2 eigenvectors of the covariance which gives the worst results.

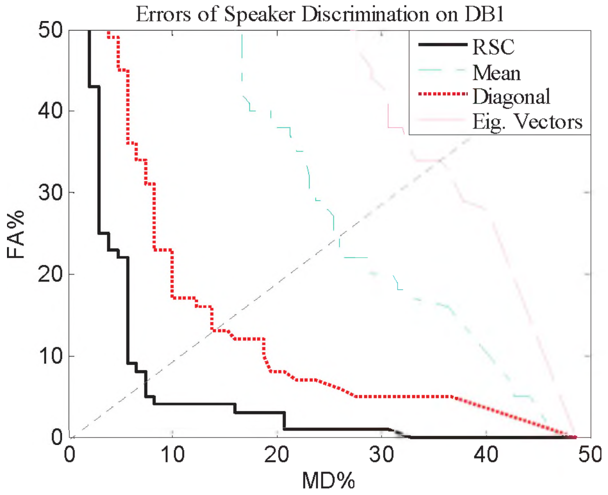


Fig. 5: Errors of speaker discrimination on DB1 - Comparison of different features: RSC, Mean of the covariance, Diagonal of the covariance and the first 2 Eigen-Vectors of the covariance.

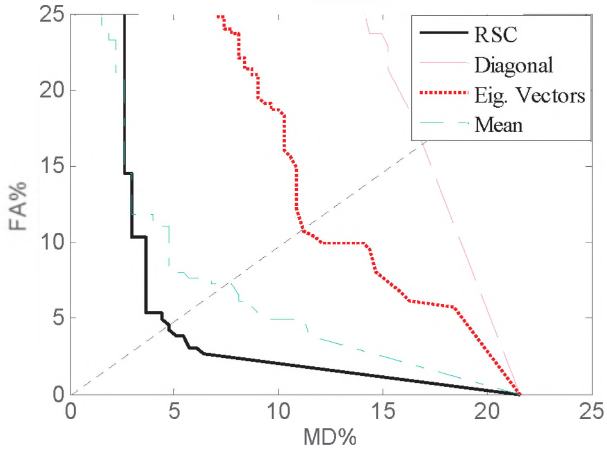


Fig. 6: Errors of speaker discrimination on TB1 - Comparison of different features: RSC, Mean of the covariance, Diagonal of the covariance and the first 2 Eigen-Vectors of the covariance.

5.3 Discriminative performance of the RSC based neural classifier

The second part of the experiment consists in comparing between the MLP-DRSC and the mono-gaussian statistical classifier. Figure 7 and figure 8 represent the ROC curves of the errors for the two different classifiers (MLP and Statistical measure) in Hub4 Broadcast-News and telephonic speech, respectively. For Hub4 Broadcast-News with segments of 4 seconds, we notice that the MLP-DRSC gives an EER of 9.25% while the EER given by the statistical measure is 11.75%. For the telephone speech with segments of 10 seconds, we notice that the MLP-DRSC gives an EER of 3.83% while the EER caused by the statistical measure is 5.74%. Therefore, the MLP-DRSC looks better than the statistical method in the two cases, especially in the medium area of the ROC curve.

Trying to further enhance the discrimination performance, one technique of fusion is proposed between the neural classifier and the statistical classifier, by using a weighted sum of the scores (Kittler, 2005) obtained by each classifier alone.

Table 3: Equal Error Rates obtained, with the different classifiers and the fusion, on different databases DB1, DB2, TB1 and TB2.

CLASSIFIER	EER % in Hub4 Broadcast-News		EER % in real Telephonic talks	
Databases	DB2	DB1	TB2	TB1
Statistical Measure	11.75	11.75	5.74	5.74
NN-DRSC	9.25	7.20	3.83	5.02
Fusion	7.88	6.77	3.65	4.29

Results of that fusion (Verlinde, 1999; Kittler, 2005), on the different databases, are shown in table 3, where it is seen that this last fusion method gives an EER better than the EER obtained by each method alone.

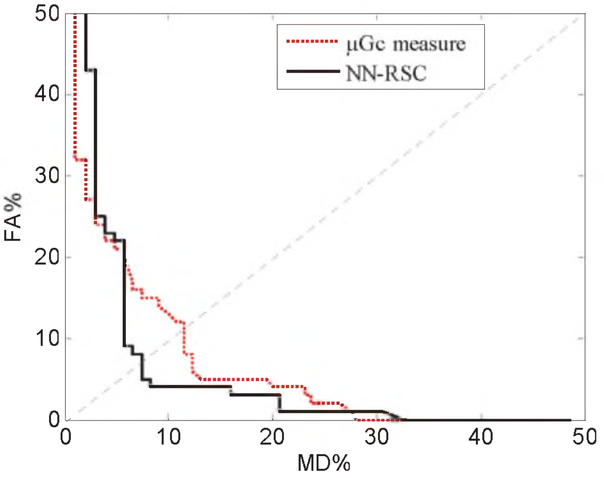


Fig. 7: Errors of speaker discrimination on DB1 (Hub4 Broadcast-News) -Comparison between the MLP-DRSC and the mono-gaussian statistical classifier.

Results presented in table 3, figure 7 and figure 8 show that the neural classifier using the relative characteristic is very interesting in speaker discrimination, on both microphonic and telephonic speech, comparatively to the statistical classifier that is evaluated in the same experimental conditions. Moreover, the fusion technique, between the two classifiers, based on the weighted sum of the scores has further improved the discrimination accuracy, where the EER is reduced in all the databases.

6. DISCUSSION AND CONCLUSION

This research work is a part of an overall project designed for audio documents indexing (Meignier, 2006), and based on speaker discrimination. However, this investigation concerns only the speaker discrimination part (Rose, 2007).

So, the major goal is to improve the discriminative performance of some existing discriminative classifiers, without altering their architecture. For that reason, we have proposed the introduction of the relativity notion in speaker modelization, by the use of a relative reduced characteristic at the input of the discriminative classifiers. We have called it: RSC or Relative Speaker Characteristic. In order to evaluate the pertinence of this new relative characteristic, two experiments were conducted:

- The first experiment was concerned with the comparison between the RSC and other existing features namely: diagonal of the covariance, mean vector and the first 2 eigenvectors of the covariance.

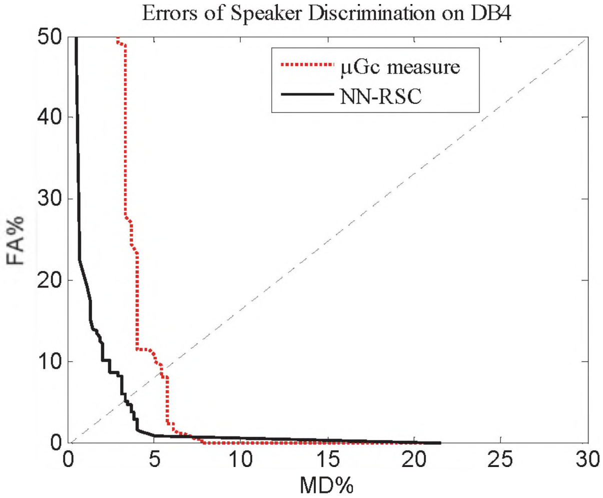


Fig. 8: Errors of speaker discrimination on TB2 (telephonic speech) - the MLP-DRSC and the mono-gaussian statistical classifier.

- The second experiment dealt with the investigation of a neural classifier using this new characteristic, in order to assess its discriminative performance with respect to a classical statistical classifier.

Discrimination experiments are done on different databases (Hub4 Broadcast-News and telephonic talks) and with different speaker modelizations. Results show that the best used modelization is based on the relative speaker characterization. This one, when used at the input of a multi-layer perceptron, provides the best scores comparatively to other types: we get an EER of 7.20% on Hub4 Broadcast-News (with segments of 4 seconds) and an EER of 3.83% on telephonic speech (with segments of 10 seconds). Thereafter, a technique of fusion was applied between the different classifiers, and experiments show that this fusion can further improve the performances.

In addition to the benefit obtained in accuracy, other benefits were noticed by using the relative characterization, such as the reduction of the training set size, reduction of the learning time and optimization of the NN convergence. Furthermore this relativity approach is really interesting due to its simplicity compared to existing techniques like PCA or LDA, and especially because it does not require any preliminary processing for the RSC estimation.

Finally, this research work shows the efficiency of the relativist approach in speaker discrimination. This new characteristic gives to the speaker a flexible model, since it changes every time that the competing speaker model changes. Although classical methods of speaker modelization consider only the speech signal of the speaker alone, the new relative modelization operates differently by using the relative speech features of the two speakers (to compare) at the input of the classifier, which is suitable in the case of speaker discrimination in difficult environments.

ACKNOWLEDGEMENTS

We would like to express our sincere acknowledgements to the Editor-In-Chief Prof. Ramani Ramakrishanan, Dr J.F. Bonastre from the LIA of Avignon and Prof. G. Cholet from the ENST of Paris for their useful helps. We also want to thank the reviewers for their constructive comments.

REFERENCES

(Atal, 1974) B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustics Society of America*, 55:1304-1312, 1974.

(Bennani, 1992) Y. Bennani. Approches connexionnistes pour la reconnaissance du locuteur: modélisation et identification. Phd thesis, Paris XI University, 1992.

(Bennani, 1995) Y. Bennani, and P. Gallinari. Neural Networks for discrimination and modelization of speakers. *Speech Communication,* volume 17, number 1-2, pp. 159-175, 1995.

(Bimbot, 1995) F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan. Second-Order Statistical Measures for text-independent Broadcaster Identification. *Speech Communication*, volume 17, number 1-2, pp. 177-192, August 1995.

(Bonastre, 1997) F. Bonastre, and L. Besacier. Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur. Actes du 4ème Congrès Français d'Acoustique, pp. 357-360, Marseille 14-18 April, 1997.

(Ferrer, 2006) L. Ferrer et al. The Contribution of Cepstral and Stylistic Features to SRI'S 2005 NIST Speaker Recognition Evaluation System ICASSP'06. , Toulouse, France, 15-19 May 2006

(Gish, 1990) H. Gish. Robust discrimination in automatic speaker identification. *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 289-292, New Mexico, April, 1990.

(Kittler, 2005) J. Kittler. Multiple classifier systems in decision-level fusion of multimodal biometric experts, *1st BioSecure residential workshop*, Paris, France 1- 26 August, 2005.

(Lee, 1995) H. S. Lee and A.C. TSOI. Application of multi-layer perceptron in estimating speech / noise characteristics for speech recognition in noisy environment. *Speech Communication*, volume. 17, number 1-2, pp. 59-76, August 1995.

(Mami, 2006) Y. Mami and D. Charlet. Speaker recognition by location in the space of reference speakers. Speech Communication 48 (2006) pp. 127 –141.

(Magrin, 2000) I. Magrin-Chagnolleau, G. Gravier, M. Seck, O. Boeffard, R. Blouet, and F. Bimbot. A Further investigation on speech features for speaker characterization. *ICSLP* 2000.

(Meignier, 2002) S. Meignier. Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection. PhD thesis, LIA Avignon, France, 2002.

(Meignier, 2006) S. Meignier et al. Step- by- step and integrated approaches in broadcast news speaker diarization. Computer Speech and Language 20 (2006) 303 –330

(Reynolds, 1995) D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, volume 17, number 1-2, pp. 91-108, 1995.

(Rose, 2007) P. Rose. Forensic Speaker Discrimination with Australian English Vowel Acoustics. ICPhS XVI Saarbrücken, 6-10 August 2007, Saarbrücken.

(Sayoud, 2000) H. Sayoud, and S. Ouamour. Reconnaissance Automatique du Locuteur en Milieu Bruité. *JEP '00*, pp. 345-348, Aussois, France, June 2000.

(Sayoud, 2003a) H. Sayoud, S. Ouamour, and M. Boudraa. 'ASTRA' An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation. *Acta Acustica*, volume 89, number 4, pp. 702-710, 2003.

(Sayoud, 2003b) H. Sayoud. Automatic speaker recognition using neural approaches. PhD thesis, USTHB University, Algiers, Algeria, 2003.

(Sayoud, 2006) H. Sayoud, and S. Ouamour. Looking for the Best Spectral Resolution in Automatic Speaker Recognition. *IEEE-GCC The 3rd Industrial Electrical & Electronics GCC Conference*, Manama, Bahrain, 19 – 22 March, 2006.

(Verlinde, 1999) P. Verlinde. A Contribution to Multimodal Identity Verification using Decision Fusion. PHD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, September.

(Wang, 2003) X. Wang, and K.K. Paliwal. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recognition 36, pp. 2429 - 2439, 2003.

(Woodland, 1997) P.C. Woodland, M.J.F. Gales, D. Pye, and S.J. Young. The Development of the 1996 HTK broadcast news transcription system. In: *DARPA Speech Recognition Workshop*, pp. 97-99, 1997.

Freedom Step

Convert a standard floor to a superior floor with the Freedom Step Acoustical & Impact Isolation Subfloor

AcoustiFloat®
Acoustical & Impact Subfloor Systems

WILREP LTD.
Tel. (905) 625-8944 Toll Free 1-888-625-8944
www.acoustifloat.com

Gym Rooms Playrooms Home Theaters Dance Floors

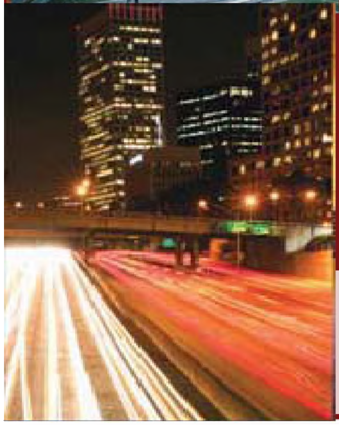
AcoustiFloat is a registered Trademark of *WILREP LTD.*

Dalimar Instruments Inc.

votre service à your service depuis since 1986

831 sound level meter/real time analyzer

- Consulting engineers
- Environmental noise monitoring
- Highway & plant perimeter noise
- Aircraft noise
- General Surveys
- Community noise

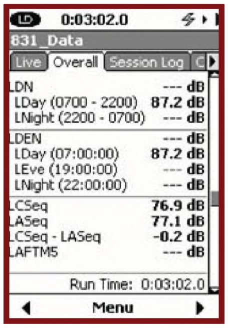


FEATURES

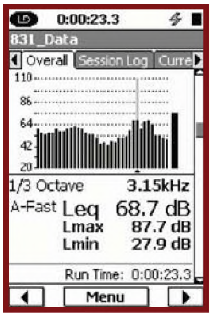
- Class 1/Type 1 sound level meter
- Small size with large display. Ergonomic
- User friendly operator interface
- 120MB standard memory expandable up to 2GB
- Single measurement range from 20 to 140 dB SPL
- Up to 16 hours of battery life
- Provided with utility software for instrument set-up and data download
- Field upgradeable
- AUX port for connection to USB mass storage & cellular modems

MEASUREMENT CAPABILITIES

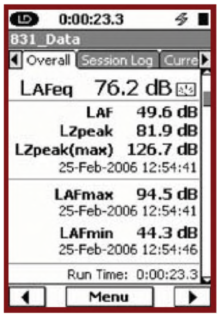
- Real time 1/1 & 1/3 octave frequency analysis
- Simultaneous display of several noise measurements—*ANY DATA* (Leq, Lmax, Spectra, etc
- Automatic logging of user selectable noise measurements (Leq, Lmax, Spectra, etc…)
- Exceedance logging with user selectable trigger levels
- Audio and voice recording with replay



```
0:03:02.0
831_Data
Live  Overall  Session Log  C
LDN                          --- dB
 LDay (0700 - 2200)     87.2 dB
 LNight (2200 - 0700)    --- dB
LDEN                         --- dB
 LDay (07:00:00)        87.2 dB
 LEve (19:00:00)         --- dB
 LNight (22:00:00)       --- dB
LCSeq                   76.9 dB
LASeq                   77.1 dB
LCSeq - LASeq           -0.2 dB
LAFTM5                   --- dB
         Run Time: 0:03:02.0
            Menu
```



```
0:00:23.3
831_Data
Overall  Session Log  Curre
1/3 Octave           3.15kHz
A-Fast Leq          68.7 dB
       Lmax         87.7 dB
       Lmin         27.9 dB
         Run Time: 0:00:23.3
            Menu
```



```
0:00:23.3
831_Data
Overall  Session Log  Curre
LAFeq          76.2 dB
            LAF      49.6 dB
         LZpeak      81.9 dB
    LZpeak(max)     126.7 dB
   25-Feb-2006 12:54:41
         LAFmax      94.5 dB
   25-Feb-2006 12:54:41
         LAFmin      44.3 dB
   25-Feb-2006 12:54:46
         Run Time: 0:00:23.3
            Menu
```

Larson Davis
A PCB Group Co.

WWW.DALIMAR.CA INFO@DALIMAR.CA
QUEBEC 450-424-0033 ONTARIO 905-707-9000 ALBERTA 403-288-4416