Second-Order Statistical Measures for Text-Independent Speaker Identification

Frédéric BIMBOT, Ivan MAGRIN-CHAGNOLLEAU and Luc MATHAN

Ecole Nationale Supérieure des Télécommunications
E.N.S.T. / Télécom Paris – Département Signal
C.N.R.S. – URA 820
46, rue Barrault
75634 PARIS cedex 13
FRANCE, European Union

E-mail: bimbot@sig.enst.fr and ivan@sig.enst.fr

Speech Communication, Vol. 17, No. 1–2, pp. 177–192, August 1995.

Second-Order Statistical Measures for Text-Independent Speaker Identification

Abstract

This article presents an overview of several measures for speaker recognition. These measures relate to second-order statistical tests, and can be expressed under a common formalism. Alternate formulations of these measures are given and their mathematical properties are studied. In their basic form, these measures are asymmetric, but they can be symmetrized in various ways. All measures are tested in the framework of text-independent closed-set speaker identification, on 3 variants of the TIMIT database (630 speakers) : TIMIT (high quality speech), FTIMIT (a restricted bandwidth version of TIMIT) and NTIMIT (telephone quality). Remarkable performances are obtained on TIMIT but the results naturally deteriorate with FTIMIT and NTIMIT. Symmetrization appears to be a factor of improvement, especially when little speech material is available. The use of some of the proposed measures as a reference benchmark to evaluate the intrinsic complexity of a given database under a given protocol is finally suggested as a conclusion to this work.

Abstandsmaße basierend auf statistischen Methoden zweiter Ordnung zur textunabhängigen Sprecheridentifizierung

Zusammenfassung

Dieser Artikel beschreibt mehrere Abstandsmaße der Sprechererkennung. Diese Abstandsmaße beziehen sich auf Tests basierend auf statistischen Methoden zweiter Ordnung und können unter einem gemeinsamen Formalismus betrachtet werden. Alternative Formalismen werden vorgestellt und ihre mathematischen Eigenschaften untersucht. In ihrer ursprünglichen Form sind diese Abstandsmaße asymetrisch. Sie können jedoch auf vielfältige Weise in eine symetrische Form umgewandelt werden. Alle Abstandsmaße werden im Rahmen einer textunabhängigen Sprechererkennung einer geschlossenen Sprechermenge an drei Variationen der TIMIT-Sprachdatenbank (630 Sprecher) getestet : TIMIT (Sprache mit hoher Aufnahmequalität), FTIMIT (eine Version von TIMIT mit eingeschränkter Bandbreite) und NTIMIT (Telephonqualität). Beachtenswerte Ergebnisse wurden mit TIMIT erreicht, die sich mit FTIMIT und NTIMIT verschlechtern. Es stellt sich heraus, daß die Symetrisierung einen Verbesserungsfaktor darstellt, vor allem, wenn wenig Sprachmaterial vorhanden ist. Die Verwendung einiger der vorgeschlagenen Abstandsmaße als Referenzvergleich zur Evaluierung der Komplexität einer gegebenen Sprachdatenbank unter einem gegebenen Protokoll wird am Ende dieser Arbeit vorgeschlagen.

Mesures statistiques du second ordre pour l'identification du locuteur indépendante du texte

Résumé

Cet article présente un ensemble de mesures pour la reconnaissance du locuteur. Ces mesures reposent sur des tests statistiques du second ordre, et peuvent être exprimées sous un formalisme commun. Différentes expressions de ces mesures sont proposées et leurs propriétés mathématiques sont étudiées. Dans leur forme la plus simple, ces mesures ne sont pas symétriques, mais elles peuvent être symétrisées de différentes façons. Toutes les mesures sont testées dans le cadre de l'identification du locuteur indépendante du texte en ensemble fermé, sur 3 versions de la base de données TIMIT (630 locuteurs) : TIMIT (parole de très bonne qualité), FTIMIT (version filtrée de TIMIT) et NTIMIT (qualité téléphonique). Des performances remarquables sont obtenues sur TIMIT, mais les résultats se dégradent naturellement avec FTIMIT et NTIMIT. La symétrisation apparaît comme un facteur d'amélioration, plus particulièrement lorsque l'on dispose de peu de parole. Il est finalement suggéré, comme conclusion à ce travail, d'utiliser certaines mesures proposées comme méthodes de référence pour évaluer la complexité intrinsèque d'une base de données quelconque, sous un protocole donné.

1 Introduction

1.1 A brief overview

Recent experiments [2] [16] [5] [17] using vector Auto-Regressive models for speaker recognition confirm and further develop work carried out by Grenier [13]. The vector AR approach provides excellent results on a subset of the TIMIT database (420 speakers), in a text-independent mode : with a training of 5 sentences (approximately 15 seconds) and tests of 1 sentence (approximately 3 seconds), closed-set identification scores reported by Montacié [17] are of 98.4 %, and reach 100 % when using 5 sentences for testing. By incorporating a discriminant analysis, the 98.4 % score improves to 99.3 %.
On the same database, other approaches have been recently tested, in particular Neural Network based methods. For instance, Rudasi and Zahorian propose binary discriminative networks [20], and reach a 100 % identification score, with 47 speakers, 5 sentences for training and 5 others for testing. Bennani [3] reports experiments with a modular TDNN-based architecture which provides 100 % correct identification for more than 100 TIMIT speakers, using about 15 seconds for training and less than 1 second for testing.
An other method used by Hattori [14] is based on predictive networks. Under this approach, a neural network is trained, for each speaker, to predict a speech frame given the 2 previous ones. During recognition, the identified speaker is the one corresponding to the network with the lowest prediction error. With the best variant, Hattori obtains 100 % correct identifications on 24 speakers (from TIMIT), with about 15 seconds for training and 9 seconds for testing.
Still on TIMIT database, Reynolds [19] shows that a Gaussian Mixture speaker model (with 32 Gaussian distributions with diagonal covariance matrices) leads to a very high level of identification performance : 99.7 % for 168 speakers, using 8 sentences for training and 2 for testing. As discussed by Furui [9], the Gaussian Mixture approach shares strong similarities with the Vector Quantization based approaches [22] and with the Ergodic HMM based methods [18] [21]. It is therefore very likely that these approaches would also provide excellent results on TIMIT.

1.2 Motivation

In spite of the fact that all approaches mentioned in this brief overview were tested on the same database, it is still difficult to have a clear idea of their relative performances. Among the factors of variability between the experiments are the speech signal pre-processing, the type of acoustic analysis, the length of training and test utterances and of course the number of speakers for which the results are reported.
A systematic comparison of any new approach with all pre-existing methods, under the exact same protocol, is theoretically possible but practically unfeasible ; not only owing to the amount of work involved, but also because it may be very difficult to reproduce in detail a specific algorithm for which all needed information may not be publicly available, or which is sensitive to initialization conditions. Moreover, it can be argued that such or such database is easy and non-discriminant (which may very well be the case for TIMIT), but we lack reliable tools to evaluate the intrinsic difficulty of a database.

A possible way to address this problem of evaluation is the use of a common algorithm as a reference benchmark to evaluate the complexity of a given database under a given protocol [7]. Desirable properties for such a reference method are its relative efficiency and robustness, but also its easy implementation and its absolute reproductibility [6].
The work reported in this article is dedicated to similarity measures between speakers which are derived from statistical tests, with an underlying Gaussian speaker model. The theoretical formulation of these measures illustrates their straightforward reproductibility, while the experimental results evaluate their efficiency on several databases : namely TIMIT (high quality speech), FTIMIT (a 0-4 kHz version of TIMIT) and NTIMIT (telephone quality speech).
In parallel to this large scale evaluation, we discuss the possibility of using one or two of the proposed approaches as systematical benchmarks, in order to provide baseline performance for any database and protocol. Such reference scores would give an idea of the degree of complexity of a given task, and the improvement obtained by any other method would indicate the benefits of a more elaborate speaker model.

1.3 Outline

Three families of measures are investigated in this paper, namely :

- log-likelihood based measures
- sphericity test based measures
- relative eigenvalue deviation measures

In section 3, we present all measures under a common formalism (defined in section 2), and we study their mathematical properties. In their original forms, these measures are not symmetric, and we describe, in section 4, some possibilities to symmetrize them. Section 5 is dedicated to the description of our evaluation protocol, and to the corresponding results. In section 6, we discuss, the possibility of using some of the measures as reference methods.

2 Notation, definitions, properties

2.1 A Gaussian model per speaker

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 a speaker $\mathcal{X}$. For instance : filter-bank coefficients, linear prediction coefficients, cepstrum coefficients,... Under the hypothesis of a Gaussian speaker model, the vector sequence $\{x_t\}$ can be summarized by its mean vector $\bar{x}$ and its covariance matrix X, i.e.

$$\bar{x} = \frac{1}{M} \sum_{t=1}^{M} x_t \quad \textit{and} \quad X = \frac{1}{M} \sum_{t=1}^{M} (x_t - \bar{x}) \cdot (x_t - \bar{x})^T \tag{1}$$

Similarly, for a speaker $\mathcal{Y}$, a parameterized speech utterance $\{y_t\}$ of N vectors can be modeled by $\bar{y}$ and Y, with

$$\bar{y} = \frac{1}{N} \sum_{t=1}^{N} y_t \quad \textit{and} \quad Y = \frac{1}{N} \sum_{t=1}^{N} (y_t - \bar{y}) \cdot (y_t - \bar{y})^T \tag{2}$$

Vectors $\bar{x}$ and $\bar{y}$ are p-dimensional, while X and Y are $p \times p$ symmetric matrices. Throughout this article, a speaker $\mathcal{X}$ (respectively $\mathcal{Y}$) will be represented by $\bar{x}$, X and M, (respectively $\bar{y}$, Y and N). We will also denote

$$\delta = \bar{y} - \bar{x}$$

$$\Gamma = X^{-\frac{1}{2}} \, Y \, X^{-\frac{1}{2}}$$

$$\rho = \frac{N}{M}$$

where $X^{\frac{1}{2}}$ is the symmetric square root matrix of X. Note that, when swapping $\mathcal{X}$ and $\mathcal{Y}$, vector δ becomes $-\delta$, matrix Γ becomes Γ^{-1} and real number ρ becomes $1/\rho$.

2.2 Second-order statistical measures

We focus on similarity measures μ between speakers $\mathcal{X}$ and $\mathcal{Y}$ which can be expressed as a function

$$\mu\,(\mathcal{X},\ \mathcal{Y}) = \phi\,(\bar{x},\ X,\ M,\ \bar{y},\ Y,\ N) \tag{3}$$

The measures μ that we investigate are derived from statistical hypothesis testing. They are constructed so that they are non-negative, i.e.

$$\forall\ \mathcal{X},\ \forall\ \mathcal{Y}, \quad \mu\,(\mathcal{X},\ \mathcal{Y})\ \geq\ 0 \tag{4}$$

and they satisfy the property

$$\forall\ \mathcal{X}, \quad \mu\,(\mathcal{X},\ \mathcal{X})\ =\ 0 \tag{5}$$

In their basic forms, the measures are non-symmetric, but we propose several ways to symmetrize them, so that

$$\forall\ \mathcal{X},\ \forall\ \mathcal{Y}, \quad \mu\,(\mathcal{X},\ \mathcal{Y})\ =\ \mu\,(\mathcal{Y},\ \mathcal{X}) \tag{6}$$

2.3 Relative eigenvalues

We will denote as $\{\lambda_i\}_{1 \leq i \leq p}$ the eigenvalues of matrix Γ, i.e. the roots of the equation

$$det\,[\Gamma\ -\ \lambda I] = 0 \tag{7}$$

where det denotes the determinant, and I the unit matrix. Matrix Γ can be decomposed as

$$\Gamma\ =\ \Theta\,\Lambda\,\Theta^{-1} \tag{8}$$

where Λ is the $p \times p$ diagonal matrix of the eigenvalues, and Θ the $p \times p$ matrix of the eigenvectors. Classically, the eigenvalues λ_i are sorted in decreasing order when i increases.
Solutions of equation (7) are known as the eigenvalues of Y relative to X. Because X and Y are positive matrices, all eigenvalues λ_i are positive. Note also that the eigenvalues of X relative to Y (i.e. the eigenvalues of Γ^{-1}) are $\{1/\lambda_i\}_{1 \leq i \leq p}$.

2.4 Mean functions of the eigenvalues

Three particular functions of the eigenvalues λ_i are used in this article :

$$\text{The arithmetic mean :} \quad a\,(\lambda_1, ..., \lambda_p) = \frac{1}{p} \sum_{i=1}^{p} \lambda_i \tag{9}$$

$$\text{The geometric mean :} \quad g\,(\lambda_1, ..., \lambda_p) = \left(\prod_{i=1}^{p} \lambda_i \right)^{1/p} \tag{10}$$

$$\text{The harmonic mean :} \quad h\,(\lambda_1, ..., \lambda_p) = \left(\frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i} \right)^{-1} \tag{11}$$

Because all eigenvalues λ_i are positive, it can be shown that

$$a \geq g \geq h \tag{12}$$

with equality if and only if all λ_i are equal. Moreover, swapping $\mathcal{X}$ and $\mathcal{Y}$ turns a into $1/h$, g into $1/g$ and h into $1/a$. In other words,

$$a\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{h\,(\lambda_1, ..., \lambda_p)} \tag{13}$$

$$g\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{g\,(\lambda_1, ..., \lambda_p)} \tag{14}$$

$$h\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{a\,(\lambda_1, ..., \lambda_p)} \tag{15}$$

2.5 Computation of a, g and h

Given that the trace (denoted tr) satisfies $tr\,(AB) = tr\,(BA)$ and that the determinant (denoted det) verifies $det\,(AB) = det\,A \cdot det\,B$, we have the following properties :

$$a\,(\lambda_1, ..., \lambda_p) = \frac{1}{p}\,tr\,\Lambda = \frac{1}{p}\,tr\,\Gamma = \frac{1}{p}\,tr\,(YX^{-1}) \tag{16}$$

$$g\,(\lambda_1, ..., \lambda_p) = (det\,\Lambda)^{1/p} = (det\,\Gamma)^{1/p} = \left(\frac{det\,Y}{det\,X}\right)^{1/p} \tag{17}$$

$$h\,(\lambda_1, ..., \lambda_p) = \frac{p}{tr\,(\Lambda^{-1})} = \frac{p}{tr\,(\Gamma^{-1})} = \frac{p}{tr\,(XY^{-1})} \tag{18}$$

These equations show that functions a, g and h can be computed directly from X, Y, X^{-1}, Y^{-1}, $det\,X$ and $det\,Y$, without extracting explicitely the eigenvalues λ_i, nor calculating the matrix square roots of X and Y. Moreover, $tr\,(YX^{-1})$ and $tr\,(XY^{-1})$ can be computed without calculating the full matrix product, but only the diagonal elements of the product.

3 Second-order statistical measures

3.1 Gaussian likelihood measure

3.1.1 Definition

By supposing that all acoustic vectors extracted from the speech signal uttered by speaker $\mathcal{X}$ are distributed like a Gaussian function, the likelihood of a single acoustic vector y_t uttered by speaker $\mathcal{Y}$ is classically

$$G\,(y_t \mid \mathcal{X}) \;=\; \frac{1}{(2\pi)^{\frac{p}{2}}\;(det\;X)^{\frac{1}{2}}}\;e^{-\frac{1}{2}(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \tag{19}$$

If we assume that all vectors y_t are independent observations, the average log-likelihood of $\{y_t\}_{1\leq t\leq N}$ can be written

$$\begin{aligned}\overline{G_{\mathcal{X}}}\,(y_1^N) \;&=\; \frac{1}{N}\,log\;G\,(y_1\;...\;y_n \mid \mathcal{X}) \;=\; \frac{1}{N}\;\sum_{t=1}^{N}\;log\;G\,(y_t \mid \mathcal{X}) \\ &=\; -\frac{1}{2}\;\left[\;p\,log\,2\pi\;+\;log\;(det\;X)\;+\;\frac{1}{N}\;\sum_{t=1}^{N}\;(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})\;\right]\end{aligned} \tag{20}$$

By replacing $y_t - \bar{x}$ by $y_t - \bar{y} + \bar{y} - \bar{x}$ and using the property

$$\frac{1}{N}\;\sum_{t=1}^{N}\;(y_t-\bar{y})^T X^{-1}(y_t-\bar{y}) \;=\; tr\;(Y X^{-1}) \tag{21}$$

we get

$$\overline{G_{\mathcal{X}}}\,(y_1^N)\;+\;\frac{p}{2}\,log\,2\pi \;=\; -\frac{1}{2}\;\left[\;log\;(det\;X)\;+\;tr\;(Y X^{-1})\;+\;(\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\;\right] \tag{22}$$

and

$$\begin{aligned}&\frac{2}{p}\;\overline{G_{\mathcal{X}}}\,(y_1^N)\;+\;log\,2\pi\;+\frac{1}{p}\;log\;(det\;Y)\;+\;1 \\ &\quad=\;\frac{1}{p}\;\left[log\left(\frac{det\,Y}{det\,X}\right)\;-\;tr\,(Y X^{-1})\;-\;(\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right]\;+\;1\end{aligned} \tag{23}$$

Therefore, if we define the Gaussian likelihood measure μ_G as

$$\mu_G\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\left[tr\,(Y X^{-1})\;-\;log\;\left(\frac{det\,Y}{det\,X}\right)\;+\;(\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right]\;-\;1 \tag{24}$$

$$=\;\frac{1}{p}\left[\;tr\;\Gamma\;-\;log\;(det\;\Gamma)\;+\;\delta^T X^{-1}\delta\;\right]\;-\;1 \tag{25}$$

$$=\;a\;-\;log\;g\;+\;\frac{1}{p}\;\delta^T X^{-1}\delta\;-\;1 \tag{26}$$

we have

$$\underset{\mathcal{X}}{Argmax}\;\;\overline{G_{\mathcal{X}}}\,(y_1^N) \;=\; \underset{\mathcal{X}}{Argmin}\;\;\mu_G\,(\mathcal{X},\,\mathcal{Y}) \tag{27}$$

3.1.2 Properties of μ_G

Matrix X^{-1} being, like X, positive definite, $\delta^T X^{-1}\delta \geq 0$. Moreover, we have $log\, g \leq g-1$ and $a \geq g$. Therefore, $a - log\, g - 1 \geq 0$ and $\mu_G(\mathcal{X}, \mathcal{Y}) \geq 0$. Measure $\mu_G(\mathcal{X}, \mathcal{Y}) = 0$ if and only if all eigenvalues λ_i are equal to 1 and δ is the null vector, i.e. if and only if $X = Y$ and $\bar{x} = \bar{y}$. However, $\mu_G(\mathcal{X}, \mathcal{Y})$ is non-symmetric, its dual term being

$$\mu_G(\mathcal{Y}, \mathcal{X}) = \frac{1}{h} + log\, g + \frac{1}{p}\,\delta^T Y^{-1}\delta - 1 \neq \mu_G(\mathcal{X}, \mathcal{Y}) \tag{28}$$

3.1.3 A variant of μ_G

When dealing with noisy or distorted speech, the mean vectors $\bar{x}$ and $\bar{y}$ may be strongly influenced by the channel characteristics, while covariance matrices X and Y are usually more robust to variations between recording conditions and transmission lines [11]. Thus, the difference $\delta = \bar{y} - \bar{x}$ may be a misleading term in μ_G.
A Gaussian likelihood measure on the covariance matrices only, denoted here μ_{Gc}, can therefore be derived from the previous likelihood measure as

$$\mu_{Gc}(\mathcal{X}, \mathcal{Y}) = \frac{1}{p}\left[\, tr\,(YX^{-1}) - log\left(\frac{det\, Y}{det\, X}\right)\right] - 1 \tag{29}$$

$$= \frac{1}{p}\,[\, tr\,\Gamma - log\,(det\,\Gamma)\,] - 1 \tag{30}$$

$$= a - log\, g - 1 \tag{31}$$

This measure can be expressed as a function of the eigenvalues λ_i of matrix Γ. However, it does not require an explicit extraction of the eigenvalues. It has the same properties as measure μ_G. In particular, it is still non-symmetric, since

$$\mu_{Gc}(\mathcal{Y}, \mathcal{X}) = \frac{1}{h} + log\, g - 1 \neq \mu_{Gc}(\mathcal{X}, \mathcal{Y}) \tag{32}$$

3.2 Arithmetic-geometric sphericity measure

3.2.1 Definition

As presented by Anderson [1], a likelihood function for testing the proportionality of a covariance matrix Y to a given covariance matrix X is

$$S(Y \mid X) = \left[\frac{det\,(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})}{\left(\frac{1}{p}\,tr\,(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})\right)^p}\right]^{\frac{N}{2}} = \left[\frac{det\,\Gamma}{\left(\frac{1}{p}\,tr\,\Gamma\right)^p}\right]^{\frac{N}{2}} \tag{33}$$

This expression results from the combination of two criteria : one on the diagonality of matrix Γ, and a second one on the equality of the diagonal elements of Γ, given that Γ is diagonal.

By denoting as $\overline{S_{\mathcal{X}}}\,(y_1^N)$ the average likelihood function for the sphericity test,

$$\overline{S_{\mathcal{X}}}\,(y_1^N) \;=\; \frac{1}{N}\,log\;S\,(Y \mid X) \tag{34}$$

and by defining

$$\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; log\;\left[\frac{\frac{1}{p}\;tr\;\Gamma}{(det\;\Gamma)^{1/p}}\right] \tag{35}$$

$$\;=\; log\;\left[\frac{\frac{1}{p}\;tr\;(Y\,X^{-1})}{\left(\frac{det\;Y}{det\;X}\right)^{1/p}}\right] \tag{36}$$

$$\;=\; log\;\left(\frac{a}{g}\right) \tag{37}$$

we have

$$\underset{\mathcal{X}}{Argmax}\;\;\overline{S_{\mathcal{X}}}\,(y_1^N) \;=\; \underset{\mathcal{X}}{Argmin}\;\;\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \tag{38}$$

Measure μ_{Sc} appears as the logarithm of the ratio of the arithmetic and the geometric means of the eigenvalues of Y relative to X. As for measure μ_{Gc}, μ_{Sc} derives from a test on the covariance matrices only. It can be expressed as a function of the eigenvalues λ_i, but it does not require the search for the eigenvalues. The use of the arithmetic-geometric sphericity test for speaker recognition was initially proposed by Grenier [12], in the framework of text-dependent experiments.

3.2.2 Properties of μ_{Sc}

Since $a \;\geq\; g$, it is obvious that $\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \;\geq\; 0$. Measure $\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; 0$ if and only if all eigenvalues λ_i are equal, i.e. if and only if X and Y are proportional. In particular, $\mu_{Sc}\,(\mathcal{X},\,\mathcal{X}) \;=\; 0$, but $X \;=\; Y$ is not a necessary condition. Finally, μ_{Sc} is not symmetric, and

$$\mu_{Sc}\,(\mathcal{Y},\,\mathcal{X}) \;=\; log\;\left(\frac{g}{h}\right) \;\neq\; \mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \tag{39}$$

3.3 Absolute deviation measure

3.3.1 Definition

The expression of μ_{Gc} and μ_{Sc} as functions of the eigenvalues λ_i are :

$$\mu_{Gc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\;\sum_{i=1}^{p}\,(\lambda_i \;-\; log\;\lambda_i \;-\; 1) \tag{40}$$

$$\mu_{Sc}\;\;(\mathcal{X},\,\mathcal{Y}) \;=\; log\;\left(\;\frac{1}{p}\;\sum_{i=1}^{p}\;\lambda_i\;\right) \;-\; \frac{1}{p}\;\sum_{i=1}^{p}\;log\;\lambda_i \tag{41}$$

As a matter of fact, it is possible to construct other metrics to measure the dissimilarity between speakers, through their covariance matrices. Any function of the eigenvalues λ_i, which is non-negative, and which takes the zero value when all eigenvalues are equal to unity, is a possible choice.
This approach was proposed by Gish [10], who constructed a measure which is based on the total absolute deviation of the eigenvalues from unity. The generic expression of this measure, which we will denote as μ_{Dc}, is

$$\mu_{Dc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\,\sum_{i=1}^{p}\;|\,\lambda_i \;-\; 1\,| \tag{42}$$

In this formulation, measure μ_{Dc} is the average absolute deviation of the eigenvalues λ_i from unity. Gish showed that robustness can be gained by removing large eigenvalues from the summation, because they may correspond to "abnormalities in small dimensional subspaces".

3.3.2 Properties of μ_{Dc}

It can be easily checked that measure μ_{Dc} is non-negative, and that it is null if and only if covariance matrices X and Y are equal. The measure is non-symmetric, since

$$\mu_{Dc}\,(\mathcal{Y},\,\mathcal{X}) \;=\; \frac{1}{p}\,\sum_{i=1}^{p}\;|\,\frac{1}{\lambda_i} \;-\; 1\,| \neq\; \mu_{Dc}\,(\mathcal{X},\,\mathcal{Y}) \tag{43}$$

4 Symmetrization

4.1 Motivation

All measures reviewed in the previous section have the common property of being non-symmetric. In other words, the roles played by the training data and by the test data are not interchangeable. However, our intuition would be that a similarity measure should be symmetric.
The asymmetry of measures μ_G, μ_{Gc} and μ_{Sc} can be explained by the following fact. These measures are based on statistical tests which suppose that the reference speaker model $\mathcal{X}$ is exact, while the test model $\mathcal{Y}$ is an estimation. But in practice, both reference and test models are estimates. Therefore, it is natural to search for a symmetric expression of originally asymmetric tests.
Moreover, it can be foreseen that the reliability of a reference model is dependent on the number of data that was used to estimate its parameters. This is experimentally confirmed by the discrepancies that can be observed in speaker identification performances, between $\mu\,(\mathcal{X},\,\mathcal{Y})$ and $\mu\,(\mathcal{Y},\,\mathcal{X})$, all the more as M and N, the number of reference and test vectors, are disproportionate (i.e. when $\rho = N/M$ is very different from 1).

4.2 Symmetrization procedures

A first possibility for symmetrizing a measure $\mu\ (\mathcal{X},\ \mathcal{Y})$, is to construct the average between the measure and its dual term :

$$\mu_{[0.5]}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\,\mu\ (\mathcal{X},\ \mathcal{Y})\ +\ \frac{1}{2}\,\mu\ (\mathcal{Y},\ \mathcal{X})\ =\ \mu_{[0.5]}\ (\mathcal{Y},\ \mathcal{X}) \tag{44}$$

For instance, the Gaussian likelihood measure, symmetrized in this manner, becomes

$$\mu_{G_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\ \left[\ a\ +\ \frac{1}{h}\ +\ \frac{1}{p}\,\delta^T\,(X^{-1}+Y^{-1})\,\delta\ \right]\ -\ 1 \tag{45}$$

which, for the covariance only measure, simplifies into

$$\mu_{Gc_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\ \left(\ a\ +\ \frac{1}{h}\right)\ -\ 1 \tag{46}$$

while the arithmetic-geometric sphericity measure becomes proportional to the arithmetic-harmonic sphericity measure [4] :

$$\mu_{Sc_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\,log\ \left(\ \frac{a}{h}\right) \tag{47}$$

This procedure of symmetrization can improve the classification performance, compared to both asymmetric terms taken individually. This is the case when training and test patterns have comparable length. However, we observed an inefficiency, or a degradation of the performance when the lengths differed significantly ($\rho\ \not\approx\ 1$). When training and test patterns are obtained from speech utterances with very different lengths, it turns out that $\mu\ (\mathcal{X},\ \mathcal{Y})$ performs better than $\mu\ (\mathcal{Y},\ \mathcal{X})$ when $\rho\ \leq\ 1$, and conversely. In other words, when the amount of training data is significantly lower than the amount of test data, it is preferable to model the test data and compute the average likelihood of the training data for the test model, rather than doing the opposite.

In the lack of a rigorous theoretical framework, we have limited our investigations to empirical trials. We have postulated an arbitrary form for more general symmetric measures μ, i.e. linear combinations of the asymmetric terms, weighted by coefficients that are function of the number of training and test vectors (respectively M and N) :

$$\mu_{[\psi_{MN}]}\ (\mathcal{X},\ \mathcal{Y})\ =\ \psi_{MN}\ \cdot\ \mu\ (\mathcal{X},\ \mathcal{Y})\ +\ \psi_{NM}\ \cdot\ \mu\ (\mathcal{Y},\ \mathcal{X}) \tag{48}$$

with

$$\psi_{MN}\ +\ \psi_{NM}\ =\ 1 \tag{49}$$

We have limited our tests to 2 particular functions ψ_{MN} and ψ_{NM}, namely :

$$\psi_{MN}\ =\ \ \alpha_{MN}\ \ =\ \frac{\sqrt{M}}{\sqrt{M}+\sqrt{N}}\ \ =\ \frac{1}{1+\sqrt{\rho}} \tag{50}$$

$$\psi_{NM}\ =\ 1-\alpha_{MN}\ =\ \frac{\sqrt{N}}{\sqrt{M}+\sqrt{N}}\ \ =\ \frac{\sqrt{\rho}}{1+\sqrt{\rho}} \tag{51}$$

and

$$\psi_{MN} \;=\; \beta_{MN} \;=\; \tfrac{M}{M+N} \;=\; \frac{1}{1+\rho} \tag{52}$$

$$\psi_{NM} \;=\; 1-\beta_{MN} \;=\; \tfrac{N}{M+N} \;=\; \frac{\rho}{1+\rho} \tag{53}$$

A similar approach was used by Montacié on AR-vector model residuals [17]. Note that, when $M \geq N$, $\rho \leq 1$ and therefore $0.5 \leq \alpha_{MN} \leq \beta_{MN}$.
We will not give the detailed expression of each measure, for each set of weights in this text. As an example, measure μ_G weighted by β_{MN} becomes

$$\mu_{G_{[\beta_{MN}]}}(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{M\,\cdot\,\mu_G(\mathcal{X},\,\mathcal{Y}) \;+\; N\,\cdot\,\mu_G(\mathcal{Y},\,\mathcal{X})}{M+N} \tag{54}$$

$$= \;\frac{1}{1+\rho}\,a \;-\; \frac{1-\rho}{1+\rho}\,log\;g \;+\; \frac{\rho}{1+\rho}\,\frac{1}{h} + \frac{1}{p}\,\delta^T\left(\frac{X^{-1}+\rho\,Y^{-1}}{1+\rho}\right)\delta \;-\; 1 \tag{55}$$

$$= \;\frac{1}{p}\left[\frac{1}{1+\rho}\,tr\,(YX^{-1}) \;-\; \frac{1-\rho}{1+\rho}\,log\left(\frac{detY}{detX}\right) \;+\; \frac{\rho}{1+\rho}\,tr\,(XY^{-1})\right] + \frac{1}{p}\left[(\bar{y}-\bar{x})^T\left(\frac{X^{-1}+\rho\,Y^{-1}}{1+\rho}\right)(\bar{y}-\bar{x})\right] \;-\; 1 \tag{56}$$

The symmetry of this expression can easily be checked.
Even though they are empirical, the symmetrizations using α_{MN} and β_{MN} provide generally better results than the symmetrization with weights equal to $\frac{1}{2}$. The optimal expression for symmetrized measures can certainly be derived from estimation theory, but it is not a trivial problem.
An exception to the general approach was applied to measure μ_{Dc}, since we experienced that it was slightly more efficient to symmetrize $log\;\mu_{Dc}$ as above, instead of μ_{Dc} itself. However, $log\;\mu_{Dc}$ can not be considered as a measure in the mathematical sense, since it is not non-negative. Therefore,

$$log\left[\mu_{Dc_{[\psi_{MN}]}}(\mathcal{X},\,\mathcal{Y})\right] \;=\; \psi_{MN}\,\cdot\,log\left[\mu_{Dc}(\mathcal{X},\,\mathcal{Y})\right] \;+\; \psi_{NM}\,\cdot\,log\left[\mu_{Dc}(\mathcal{Y},\,\mathcal{X})\right] \tag{57}$$

which is equivalent to

$$\mu_{Dc_{[\psi_{MN}]}}(\mathcal{X},\,\mathcal{Y}) \;=\; \mu_{Dc}(\mathcal{X},\,\mathcal{Y})^{\,\psi_{MN}} \;\cdot\; \mu_{Dc}(\mathcal{Y},\,\mathcal{X})^{\,\psi_{NM}} \tag{58}$$

5 Experiments and results

5.1 Task

We have tested the measures described in this article, in the framework of closed-set text-independent speaker identification. There is a single reference per speaker (composed of a mean vector $\bar{x}$, a covariance matrix X and a number of data M). All test utterances are different from all training utterances, and all training utterances are different from one another. Each measure is evaluated as regards its classification ability using a 1-nearest neighbour decision rule. The possibility of rejection is not taken into account : the test speaker is always part of the set of references.

5.2 Databases

For our experiments, we used TIMIT and NTIMIT databases. TIMIT [8] contains 630 speakers (438 male and 192 female), each of them having uttered 10 sentences. Two sentences have the prefix "sa" (sa1 and sa2). Sentences sa1 and sa2 are different, but they are the same across speakers. Three sentences have the prefix "si" and five have the prefix "sx". These 8 sentences are different from one another, and different across speakers. Sentences "sa" and "si" have an average duration of 2.9 seconds. Sentences "sx" have an average duration of 3.2 seconds. The speech signal is recorded through a high quality microphone, in a very quiet environment, with a 0-8 kHz bandwidth. The signal is sampled at 16 kHz, on 16 bits, on a linear amplitude scale. Moreover, all recordings took place in a single session (contemporaneous speech).
The NTIMIT database [15] was obtained by playing TIMIT speech signal through an artificial mouth installed in front of the microphone of a fixed handset frame and transmitting this input signal through a different telephone line for each sentence (local or long distance network). The signal is sampled at 16 kHz, but its useful bandwidth is limited to telephone bandwidth (approximately 300-3400 Hz). Each sample is represented on 16 bits (linear).

5.3 Signal analysis

Each sentence is analysed as followed : the speech signal is decomposed in frames of 504 samples (31.5 ms) at a frame rate of 160 samples (10 ms). A Hamming window is applied to each frame. The signal is not pre-emphasized. For each frame, a Winograd Fourier Transform is computed and provides 252 square module values representing the short term power spectrum in the 0-8 kHz band.
This Fourier power spectrum is then used to compute 24 filter bank coefficients. Each filter is triangular (except the first and last ones which have a rectangle trapezoidal shape). They are placed on a non-uniform frequency scale, similar to the Bark/Mel scale. The central frequency of the 24 filters are, in Hz : 47, 147, 257, 378, 510, 655, 813, 987, 1178, 1386, 1615, 1866, 2141, 2442, 2772, 3133, 3529, 3964, 4440, 4961, 5533, 6159, 6845, and 7597. Each filter covers a spectral range from the central frequency of the previous filter to the central frequency of the

next filter, with a maximum value of 1 for its own central frequency. For each frequency, only 2 filters (maximum) are non-zero, and their magnitudes add up to 1.
We finally take the base 10 logarithm of each filter output and multiply the result by 10, to form a 24-dimensional vector of filter bank coefficients in dB. For the TIMIT database, all 24 coefficients are kept, from which we compute, for each utterance a 24-dimensional mean vector and a 24×24 (symmetric) covariance matrix.
In order to simulate, for some of the experiments, a low-pass filtering of the speech signal in the 0-4 kHz band, we have simply discarded the last 7 coefficients of the 24-dimensional vectors obtained from the full band signal. The last filter, with index 17, has a central frequency of 3529 Hz, and becomes zero above 3964 Hz. This is the approach we used for NTIMIT database, since the useful bandwidth does not exceed 4000 Hz for these data. We also used this approach on TIMIT, in order to obtain results corresponding to a 0-4 kHz bandwidth, without the telephone line variability. We will refer to these data as FTIMIT data. Under these analysis conditions, each mean vector is 17 dimensional, while covariance matrices are 17×17 (symmetric) matrices.

5.4 Training and test protocols

We use 2 training protocols, namely a "long training" and a "short training".

- For the **"long training"**, we use all 5 "sx" sentences concatenated together as a single reference pattern for each speaker. The average total duration of a "long training" pattern is **14.4 seconds**. A single reference (mean vector $\bar{x}$, covariance matrix X and number of vectors M) is computed for each speaker from all speech frames, represented as filter bank coefficients. In particular, no speech activity detector is used to remove silent speech portions.

- For the **"short training"**, we only use the first 2 "sx" sentences in alphanumeric order, in the same way as for the "long training". The average total duration of a "short training" is **5.7 seconds** (including silences).

For the tests, we also have 2 distinct protocols : a "long test" and a "short test".

- For the **"long test"**, all "sa" and "si" sentences (5 in total) are concatenated together as a single test pattern, for each speaker. In this framework, we therefore have a single test pattern per speaker, i.e. **630 test patterns** altogether. In average, each pattern lasts **15.9 seconds**.

- For the **"short test"**, each "sa" and "si" sentences are tested separately. The whole test set thus consists of $630 \times 5 =$ **3150 test patterns**, of **3.2 seconds** each, in average.

Even though the "sa" sentences are the same for each speaker, these sentences are used in the test set. Therefore, the experiments can be considered as totally text-independent.

5.5 Experiments

In the experiments reported in this article, we have systematically tested the 4 families of measures :

- μ_G
- μ_{Gc}
- μ_{Sc}
- μ_{Dc}

in two asymmetric forms :

- $\mu\ (\mathcal{X},\ \mathcal{Y})$
- $\mu\ (\mathcal{Y},\ \mathcal{X})$

as well as in the three symmetric forms proposed in section 4 :

- $\mu_{[0.5]}\ (\mathcal{X},\ \mathcal{Y})$
- $\mu_{[\alpha_{MN}]}\ (\mathcal{X},\ \mathcal{Y})$
- $\mu_{[\beta_{MN}]}\ (\mathcal{X},\ \mathcal{Y})$

These evaluations were carried out on :

- TIMIT (24 filter bank coefficients between 0 and 8000 Hz)
- FTIMIT (first 17 filter bank coefficients from TIMIT between 0 and 4000 Hz)
- NTIMIT (17 filter bank coefficients between 0 and 4000 Hz)

It is reasonable to expect that the 3 databases, in this order, correspond to an increasing degree of difficulty.
In each case, we give the results for 4 possible training × test protocols, corresponding to various typical values of the total amount of speech material per speaker $T = M + N$, of the ratio ρ between test and training material, and therefore to different weighting factors α_{MN} and β_{MN} :

- **"long-long"** protocol : long training × long test
  $\bar{T} \approx 3000\, cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$
- **"short-long"** protocol : short training × long test
  $\bar{T} \approx 2150\, cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$
- **"long-short"** protocol : long training × short test
  $\bar{T} \approx 1750\, cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$
- **"short-short"** protocol : short training × short test
  $\bar{T} \approx\ \ 900\, cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

5.6 Results

The results are organized in 3 sets of 4 tables. The first set of tables (numbered I.1, I.2, I.3 and I.4) corresponds to results for TIMIT, the second set (Tables II.1 to II.4) for FTIMIT and the third set (Tables III.1 to III.4) for NTIMIT. The first table of each set (i.e. Tables I.1, II.1 and III.1) reports the results obtained for the "long-long" protocol, while the second one (I.2, II.2 and III.2) reports those for the "short-long" protocol. Similarly, the third and fourth tables of each set correspond respectively to the "long-short" and "short-short" protocols. In each table, the results relative to a given family of measures are organized in columns. The first line corresponds to the scores of both asymmetric terms (each cell is subdivided into 2), while the second, third and fourth lines show the results for the various symmetric forms. All results are given in terms of percentage of correct identification. Depending on this percentage S, and on the number of test patterns n, we give in Table 0 the half-width of the 95 % confidence interval, which is calculated as :

$$\pm 2 \sqrt{\frac{S \cdot (100 - S)}{n}}$$

Note that this quantity is the same for a score S and for $100 - S$.

score : S | 100 - S	95 | 5	85 | 15	75 | 25	65 | 35	55 | 45
long test, n = 630	± 1.7 %	± 2.8 %	± 3.5 %	± 3.8 %	± 4.0%
short test, n = 3150	± 0.8 %	± 1.3 %	± 1.5 %	± 1.7 %	± 1.8%

Table 0 : *Half-width of the 95 % confidence interval for different values of the identification score S in %, corresponding to the long and short test protocols.*

We will not comment in detail each performance figure in Tables I, II and III, but we will rather try to underline several global trends.
For all measures, $\mu\ (\mathcal{X},\ \mathcal{Y})$ and $\mu\ (\mathcal{Y},\ \mathcal{X})$ perform differently. The term $\mu\ (\mathcal{X},\ \mathcal{Y})$ performs better when the training speech material has a longer duration than the test material, and conversely. The discrepancy between the performances of the asymmetric terms is especially obvious for measure μ_{Dc}
With non-distorted speech (TIMIT and FTIMIT), measure μ_G outperforms measure μ_{Gc} and all other measures on covariance matrices only. On the opposite, when channel variability is present (NTIMIT), the use of the mean vectors is, as expected, detrimental to the results.
In their asymmetric forms, the most efficient measure among the covariance-only measures is measure μ_{Sc}. However, when symmetrisation is applied, the performances tend to level-off, with a slight advantage for μ_{Dc}.
Among the symmetrization procedures that we tested, the most efficient one seems to be the one using weights β_{MN} and β_{NM} for μ_G, μ_{Gc} and μ_{Sc}, whereas α_{MN} and α_{NM} appear to be preferable for $log\ \mu_{Dc}$.

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	100 % | 100 %	100 % | 99.8 %	100 % | 100 %	99.5 % | 99.8 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	100 %	100 %	100 %	100 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	100 %	100 %	100 %	100 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	100 %	100 %	100 %	100 %

Table I.1: *long training (5 sentences $\approx$ 14.4 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 3000\ cs,\ \bar{\rho} \approx 1.10,\ \bar{\alpha}_{MN} \approx 0.48,\ \bar{\beta}_{MN} \approx 0.49$

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	93.2 % | 99.4 %	86.7 % | 97.1 %	94.9 % | 96.4 %	73.3 % | 92.1 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	98.1 %	94.6 %	95.7 %	95.1 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	98.7 %	95.7 %	96.2 %	97.0 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	99.2 %	96.5 %	96.0 %	97.0 %

Table I.2: *short training (2 sentences $\approx$ 5.7 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 2150\ cs,\ \bar{\rho} \approx 2.79,\ \bar{\alpha}_{MN} \approx 0.26,\ \bar{\beta}_{MN} \approx 0.37$

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	97.9 % | 89.7 %	96.2 % | 78.8 %	97.3 % | 93.6 %	83.6 % | 59.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	97.2 %	93.9 %	97.0 %	97.3 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	98.4 %	97.1 %	97.3 %	97.6 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	98.4 %	97.6 %	97.6 %	94.8 %

Table I.3: *long training (5 sentences $\approx$ 14.4 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx 1750\ cs,\ \bar{\rho} \approx 0.22,\ \bar{\alpha}_{MN} \approx 0.82,\ \bar{\beta}_{MN} \approx 0.68$

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	83.8 % | 78.2 %	73.5 % | 64.9 %	81.9 % | 77.7 %	52.9 % | 45.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	89.7 %	82.2 %	82.7 %	84.4 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	90.1 %	83.4 %	83.0 %	84.2 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	89.7 %	83.6 %	83.3 %	80.1 %

Table I.4: *short training (2 sentences $\approx$ 5.7 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx\ \ 900\ cs,\ \bar{\rho} \approx 0.56,\ \bar{\alpha}_{MN} \approx 0.64,\ \bar{\beta}_{MN} \approx 0.57$

Tables I.1, I.2, I.3, I.4 :
Text-independent speaker identification – TIMIT database (630 speakers).
The results are given in percentage of correct identification.

Measures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	98.4 %	99.4 %	96.1 %	98.3 %	97.6 %	98.1 %	90.5 %	95.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		99.4 %		97.9 %		97.9 %		98.3 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		99.5 %		97.9 %		97.9 %		98.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		99.5 %		97.9 %		97.8 %		98.4 %	

Table II.1: *long training (5 sentences $\approx$ 14.4 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 3000\ cs,\ \bar{\rho} \approx 1.10,\ \bar{\alpha}_{MN} \approx 0.48,\ \bar{\beta}_{MN} \approx 0.49$

Measures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	78.7%	88.6%	63.2 %	77.0 %	72.9 %	76.4 %	44.1 %	65.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		87.0 %		76.8 %		76.4 %		76.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		87.9 %		77.5 %		76.2 %		77.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		89.0 %		77.8 %		76.4 %		76.2 %	

Table II.2: *short training (2 sentences $\approx$ 5.7 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 2150\ cs,\ \bar{\rho} \approx 2.79,\ \bar{\alpha}_{MN} \approx 0.26,\ \bar{\beta}_{MN} \approx 0.37$

Measures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	81.4 %	67.9 %	70.0 %	49.8 %	70.7 %	66.3 %	48.1 %	33.3 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		81.8 %		67.3 %		70.4 %		72.2 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		84.2 %		71.8 %		71.7 %		73.1 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		83.6 %		72.6 %		72.0 %		64.4 %	

Table II.3: *long training (5 sentences $\approx$ 14.4 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx 1750\ cs,\ \bar{\rho} \approx 0.22,\ \bar{\alpha}_{MN} \approx 0.82,\ \bar{\beta}_{MN} \approx 0.68$

Measures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	54.7%	49.7%	39.8 %	32.2 %	42.6 %	41.2 %	23.1 %	20.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		61.4 %		43.9 %		44.4 %		46.5 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		61.8 %		45.3 %		44.5 %		46.8 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		61.4 %		45.8 %		44.4 %		43.6 %	

Table II.4: *short training (2 sentences $\approx$ 5.7 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx\ \ 900\ cs,\ \bar{\rho} \approx 0.56,\ \bar{\alpha}_{MN} \approx 0.64,\ \bar{\beta}_{MN} \approx 0.57$

Tables II.1, II.2, II.3, II.4 :
Text-independent speaker identification – FTIMIT database (630 speakers).
The results are given in percentage of correct identification.

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	45.3 % | 50.3 %	59.5 % | 63.0 %	66.0 % | 64.9 %	41.0 % | 51.0 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	49.4 %	63.0 %	66.4 %	67.9 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	49.0 %	63.0 %	66.5 %	68.4 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	49.4 %	63.6 %	66.5 %	68.6 %

Table III.1: *long training (5 sentences $\approx$ 14.4 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 3000\ cs,\ \bar{\rho} \approx 1.10,\ \bar{\alpha}_{MN} \approx 0.48,\ \bar{\beta}_{MN} \approx 0.49$

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	17.6 % | 24.9 %	22.2 % | 31.0 %	28.4 % | 29.7 %	12.4 % | 22.5 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	24.4 %	29.5 %	29.8 %	30.3 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	24.8 %	30.5 %	30.0 %	30.8 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	25.7 %	31.3 %	30.5 %	30.8 %

Table III.2: *short training (2 sentences $\approx$ 5.7 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 2150\ cs,\ \bar{\rho} \approx 2.79,\ \bar{\alpha}_{MN} \approx 0.26,\ \bar{\beta}_{MN} \approx 0.37$

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	20.7 % | 13.8 %	25.4 % | 13.5 %	26.3 % | 23.0 %	14.1 % | 5.2 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	19.3 %	23.4 %	25.2 %	25.2 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	21.1 %	25.4 %	26.1 %	26.4 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	21.4 %	26.1 %	26.7 %	20.5 %

Table III.3: *long training (5 sentences $\approx$ 14.4 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx 1750\ cs,\ \bar{\rho} \approx 0.22,\ \bar{\alpha}_{MN} \approx 0.82,\ \bar{\beta}_{MN} \approx 0.68$

Measures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	10.1 % | 9.0 %	12.0 % | 8.9 %	13.7 % | 12.8 %	6.4 % | 3.1 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	11.7 %	13.7 %	14.3 %	14.4 %
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	11.7 %	14.2 %	14.4 %	14.8 %
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	11.6 %	15.0 %	14.4 %	12.7 %

Table III.4: *short training (2 sentences $\approx$ 5.7 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx\ \ 900\ cs,\ \bar{\rho} \approx 0.56,\ \bar{\alpha}_{MN} \approx 0.64,\ \bar{\beta}_{MN} \approx 0.57$

Tables III.1, III.2, III.3, III.4 :
Text-independent speaker identification – NTIMIT database (630 speakers).
The results are given in percentage of correct identification.

The positive effect of symmetrization is important when little speech material is available. The most significant differences are observed for the short training × short test protocol. Table IV gives orders of magnitude of the relative error rate reduction between the asymmetric measures and their best symmetric version. If S is the percentage of correct identification for the asymmetric measure and S' is the percentage of correct identification for the symmetric measure, the relative error rate reduction is calculated as :

$$\frac{S' - S}{100 - S}$$

This relative improvement is given for the two protocols using short duration test data only. For the two others, the statistical significance of the observed differences are too small to be conclusive, given the larger confidence interval.

measure	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
TIMIT	~ 30 %	~ 40 %	~ 10 %	~ 75 %
FTIMIT	~ 15 %	~ 10 %	~ 5 %	~ 40 %
NTIMIT	~ 1 %	~ 1 %	< 1 %	~ 10 %

Table IV : *Order of magnitude of the relative error rate reduction between asymmetric and symmetric measures. Results for short test protocols only.*

These results show that symmetrization improves covariance-only measures (μ_{Gc}, μ_{Sc} and μ_{Dc}) as the task becomes intrinsically less difficult (TIMIT > FTIMIT > NTIMIT), and as the original asymmetric measures perform less well ($\mu_{Dc} < \mu_{Gc} < \mu_{Sc}$). On the other hand, when the Gaussian speaker model is not powerful enough for the task (NTIMIT), or when the asymmetric measure is quite efficient (μ_{Sc}), symmetrization is less useful.

6 Discussion

Our evaluations show that remarkable performances can be obtained on the TIMIT database for text-independent closed-set speaker identification (630 speakers) by second-order statistical measures, i.e. with a very simple underlying speaker model. Therefore, TIMIT is certainly an easy database for speaker recognition, and the measures exposed in this article work very well, on this database. Naturally, their overall performances degrade with more adverse conditions : a significant amount of speaker characteristics seems to be contained in the 4–8 kHz band, since FTIMIT results are significantly worse than TIMIT results. The effect of telephone channel distortion and variability are the cause of an even more severe drop on NTIMIT recognition scores. The effect of temporal drift owed to multisession recordings can not be studied with TIMIT derived data, but it is easy to predict an additional negative role of this factor on the performances. If second-order statistical measures are clearly efficient for relatively simple tasks, they are obviously not the ultimate solution to speaker recognition for any kind of applications.

6.1 Beyond the performances

However, second-order statistical measures have several advantages. They are simple to implement and easy to reproduce. Moreover, Gaussian likelihood measures (μ_G and μ_{Gc}) in their asymmetric forms are particular cases of several general approaches frequently used in text-independent speaker recognition. A 1-Gaussian speaker model is equivalent to a Vector Quantization codebook with 1 entry associated with a Mahalanobis distance. It is also equivalent to any kind of Hidden Markov Model (Left-to-Right, Ergodic,...) with 1 state and 1 Gaussian distribution. It is a particular case of a k-Gaussian Mixture model with k = 1. Finally, the likelihood criterion is often used on vector prediction residuals obtained from linear or connectionist models for which the identity model (0^{th}-order prediction) is a particular case.
Therefore, μ_G and μ_{Gc} are at the intersection of several classical approaches, which are extensions of this basic model in various directions (variations of the distance measure, use of more or less strong temporal constraints, refinement of the speaker distribution model, filtering of the acoustic parameters,...). Given the extreme simplicity of the second-order statistical measures, we therefore suggest that any speaker recognition task could be systematically benchmarked by one or two of these measures, in order to obtain a reference score indicating the intrinsic complexity of the chosen database and protocol. In particular, the preprocessings, the acoustic analysis, the training and test splitting of the data, and the decision strategy for the method under test should be identically used for the benchmark method.

6.2 A possible reference approach

Even though asymmetric Gaussian likelihood based measures do not systematically perform better than other second-order statistical measures, μ_G and μ_{Gc} may be preferable as reference benchmark measures in two cases : when they are compared with other asymmetric approaches (which is the case for VQ, HMM and Gaussian Mixtures), and when the length of training material is significantly higher than the length of test material. The choice between μ_G and μ_{Gc} should be guided by the processing that is applied to the data for the system under evaluation : whether, for this particular protocol, the long term average is substracted or not to the acoustic parameters. Measures $\mu_{G_{[\beta_{MN}]}}$ or $\mu_{Gc_{[\beta_{MN}]}}$ can also be implemented simply and could be systematically tested. However, the lack of theoretical justification for these measures, and the relatively small improvement they provide as soon as a reasonable amount of speech material is available, make it more debatable. Nevertheless, if the approach under test is formally symmetric, it would be fair to compare it to a symmetric reference measure.

7 Conclusion

The goal of this work has been multiple. Firstly, to investigate the properties and performances of simple speaker recognition approaches, to compare them and to identify their limits. Our large scale evaluation on TIMIT, on a low-pass filtered version of TIMIT and on NTIMIT illustrates clearly that speech quality and quantity are major factors of performance, and that on high quality contemporaneous speech, simple and fast methods can be extremely efficient. For instance, this type of approach may prove sufficient for applications such as the automatic speaker labeling of radio or television recordings, for which the signal quality is constant and the voice drift relatively marginal.
Secondly, our work illustrates the extreme caution with which any conclusion can be drawn on the merit of a given method outside of any point of comparison. Since it may not be feasible to compare any new method with all state-of-the-art approaches, it is at least desirable to benchmark the task with a simple and general reference approach.
Thirdly, we believe that second-order statistical tests and measures, based on the Gaussian likelihood scoring are a good compromise as reference measures, since they are easy to implement, simple to reproduce, inexpensive in computation and light in storage requirements. Moreover, they appear, in their asymmetric forms, as simpler versions of more elaborate approaches. Though symmetrization is not a systematic factor of improvement, symmetric versions of the measures could be tested as well, especially in comparison with other symmetric methods. More theoretical work on symmetrization is however needed to find optimal symmetric forms.
The systematical use of a reference method in order to calibrate the complexity of a speaker recognition task can only result from a consensus between researchers both on the concept of a benchmark evaluation by a common approach and on the choice of the reference algorithm itself. We hope that this article will contribute to widen the concertation that had started during the SAM-A European ESPRIT project, dedicated to speech assessment methodology.

References

[1] T. W. Anderson. *An Introduction to Multivariate Analysis*. John Wiley and Sons, 1958.

[2] T. Artières, Y. Bennani, P. Gallinari, and C. Montacié. Connectionnist and conventional models for text-free talker identification tasks. In *Proceedings of NEURONIMES 91*, 1991. Nîmes, France.

[3] Y. Bennani. Speaker identification through a modular connectionist architecture: evaluation on the TIMIT database. In *Proceedings of ICSLP 92*, volume 1, pages 607–610, Oct. 1992. Banff, Canada.

[4] F. Bimbot and L. Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Proceedings of EUROSPEECH 93*, volume 1, pages 169–172, Sept. 1993. Berlin, Germany.

[5] F. Bimbot, L. Mathan, A. de Lima, and G. Chollet. Standard and target-driven AR-vector models for speech analysis and speaker recognition. In *Proceedings of ICASSP 92*, volume 2, pages II.5–II.8, Mar. 1992. San Francisco, United-States.

[6] F. Bimbot, A. Paoloni, and G. Chollet. *Assessment Methodology for Speaker Identification and Verification Systems*. Technical report – Task 2500 – Report I9, SAM-A ESPRIT Project 6819, 1993.

[7] G. Chollet and C. Gagnoulet. On the evaluation of speech recognizers and data bases using a reference system. In *Proceedings of ICASSP 82*, volume 3, pages 2026–2029, May 1982. Paris, France.

[8] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA speech recognition research database : specifications and status. In *Proceedings of the DARPA workshop on speech recognition*, pages 93–99, Feb. 1986.

[9] S. Furui. An overview of speaker recognition technology. In *Workshop on automatic speaker recognition, identification and verification*, pages 1–9, Apr. 1994. Martigny, Switzerland.

[10] H. Gish. Robust discrimination in automatic speaker identification. In *Proceedings of ICASSP 90*, volume 1, pages 289–292, Apr. 1990. New Mexico, United-States.

[11] H. Gish, M. Krasner, W. Russell, and J. Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *Proceedings of ICASSP 86*, volume 2, pages 865–868, Apr. 1986. Tokyo, Japan.

[12] Y. Grenier. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*. PhD thesis, ENST, 1977.

[13] Y. Grenier. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. In *XIèmes Journées d'Etude sur la Parole*, pages 163–171, May 1980. Strasbourg, France.

[14] H. Hattori. Text-independent speaker recognition using neural networks. In *Proceedings of ICASSP 92*, volume 2, pages 153–156, Mar. 1992. San Francisco, United-States.

[15] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of ICASSP 90*, Apr. 1990. New Mexico, United-States.

[16] C. Montacié, P. Deléglise, F. Bimbot, and M.-J. Caraty. Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction. In *Proceedings of ICASSP 92*, volume 1, pages 153–156, Mar. 1992. San Francisco, United-States.

[17] C. Montacié and J.-L. Le Floch. AR-vector models for free-text speaker recognition. In *Proceedings of ICSLP 92*, volume 1, pages 611–614, Oct. 1992. Banff, Canada.

[18] A. B. Poritz. Linear predictive Hidden Markov Models and the speech signal. In *Proceedings of ICASSP 82*, pages 1291–1294, May 1982. Paris, France.

[19] D. A. Reynolds. Speaker identification and verification using Gaussian Mixture speaker models. In *Workshop on automatic speaker recognition, identification and verification*, pages 27–30, Apr. 1994. Martigny, Switzerland.

[20] L. Rudasi and S. A. Zahorian. Text-independent talker identification with neural networks. In *Proceedings of ICASSP 91*, volume 1, pages 389–392, 1991. Toronto, Canada.

[21] M. Savic and S. K. Gupta. Variable parameter speaker verification system based on hidden markov modeling. In *Proceedings of ICASSP 90*, volume 1, pages 281–284, Apr. 1990. Albuquerque, New Mexico, United-States.

[22] F. K. Soong, A. E. Rosenberg, and B.-H. Juang. A Vector Quantization approach to speaker recognition. Technical Report 3, AT&T, Mar. 1987.