# Reconnaissance de locuteurs indépendante du texte

Philippe Thévenaz

Cet article décrit une technique de reconnaissance automatique de locuteurs. La particularité de cette technique est l'indépendance vis-à-vis du texte, ce qui signifie qu'un utilisateur n'a plus à se souvenir d'un mot de passe, car un locuteur peut être reconnu sur la base de n'importe quel texte. A cet effet, nous avons choisi d'utiliser quatre méthodes différentes que nous analysons tout d'abord séparément. Nous montrons ensuite comment les combiner pour améliorer la performance globale. La première méthode caractérise un locuteur par son cepstre moyen. La deuxième fait aussi usage de paramètres cepstraux; elle mesure la disparité entre deux locuteurs par une technique d'accumulation d'erreur de quantification vectorielle. La troisième méthode est presque identique à la précédente; elle en diffère par l'utilisation de la dérivée temporelle des vecteurs cepstraux, en lieu et place de leur valeur instantanée. La quatrième méthode exploite l'histogramme des vecteurs choisis par quantification vectorielle dans un vocabulaire universel. Enfin, nous combinons les résultats par la technique du discriminant linéaire de Fisher. Nous montrons, dans une série d'expériences menées en largeur de bande téléphonique, que trois de ces méthodes produisent de bons résultats, et que leur usage conjoint permet encore de sensiblement améliorer la performance globale.

## Text independent speaker recognition

This article describes an automatic speaker recognition technique, in a text independent context. By text independence, we mean that no password has to be used, reducing the risk of loss or mimicry. To this end, we select four different methods whose performances are individually reviewed. Then, we show how to combine them in order to enhance the global result. The first method characterizes a speaker by his mean cepstrum, averaged over time. The second method is based on a technique of accumulation of vector quantization error; the parameters used are also cepstral vectors. Using their time derivatives instead, we produce the third method, which is otherwise identical to the previous one. The fourth and last method exploits the histogram of entries in a universal cepstrum codebook, according to a vector quantization technique. Finally, we combine the results by the Fisher linear discriminant analysis. We show, by a series of telephone-bandwidth experiments, that the methods behave well; the third method only has to be rejected, owing to its bad performance. However, the combination of the other three methods is even more successful than any single method.

## Textunabhängige Sprechererkennung

Der vorliegende Artikel beschreibt eine Technik zur automatischen Erkennung eines Sprechers. Die Technik ist textunabhängig, was den Verzicht auf spezielle Textsequenzen wie Passwörter erlaubt. Wir haben dazu vier verschiedene Methoden ausgewählt, die zuerst einzeln betrachtet werden. Anschliessend wird untersucht, in welcher Weise diese Methoden zu kombinieren sind, um das globale Resultat zu verbessern. Die erste Methode charakterisiert den Sprecher durch den zeitlichen Mittelwert seines Cepstrums. Die zweite Methode bestimmt die Distanz zwischen zwei Sprechern mit einer Technik, die auf der Akkumulation von Vektorquantifikationsfehlern beruht. Die dritte Methode ist der vorhergehenden sehr ähnlich, nur werden die Cepstrum-Parameter zuerst differenziert. Die vierte Methode benützt das Histogramm der Eingänge in ein universelles Cepstrum-Kodebuch, das durch vektorielle Quantifikation aus einem universellen Vokabular einer grossen Population bestimmt wurde. Schlussendlich kombinieren wir die Resultate mit der linearen Diskriminationsmethode von Fisher. Wir zeigen anhand von Experimenten, die im Telefonsprachband durchgeführt wurden, dass die Methoden, ausser der dritten, gut geeignet sind, um Sprecherverifikation zu führen. Die Kombination von drei der verschiedenen Methoden erbringt sogar weit bessere Resultate als jede der Methoden einzeln.

# 1. Introduction

Cet article traite de reconnaissance de locuteurs, domaine qui consiste à faire exécuter automatiquement à une machine la tâche de déterminer l'identité d'une personne sur la seule base de sa parole [1, 4]. L'être humain est capable de résoudre quotidiennement ce problème de façon efficace, ce qui démontre l'existence d'une solution; cependant, il utilise dans cette tâche une certaine quantité d'informations qui paraissent, de nos jours encore, inaccessibles à une machine, telles que par exemple le contenu du message, d'où est extrait un contexte sémantique apte à faciliter le travail de reconnaissance. Par contraste, les outils et algorithmes utilisés dans le domaine du traitement de signal permettent d'extraire les caractéristiques acoustiques d'un morceau de parole de manière bien plus fine que ne saurait le faire un être humain. Ceci explique le succès des méthodes faisant usage d'un mot de passe, car il est possible d'y comparer une locution de test avec une locution de référence, obtenue par apprentissage, sur la base de critères purement acoustiques et temporels [3].
Le problème auquel nous nous attaquons ici est celui de reconnaître un locuteur de façon indépendante du texte qu'il prononce. Une des applications possibles est par exemple le contrôle continu de l'identité d'une personne menant une transaction téléphonique.

### 1.1 But

Notre but est de construire un système de reconnaissance du locuteur indépendant du texte et combinant une multitude de méthodes différentes [15, 19]. A cet effet, nous nous sommes concentrés sur des aspects vocaux déjà reconnus comme efficaces en tant que base à une méthode de reconnaissance du locuteur [6, 18, 23, 24, 27]. L'enjeu du choix de ces méthodes n'est pas leur performance propre, qui est déjà abondamment décrite dans la littérature, mais bien plutôt leur performance relative et conjuguée. Ainsi, nous espérons pouvoir abaisser au mieux les taux d'erreurs et minimiser les contraintes imposées à un utilisateur du système.

### 1.2 Indépendance du texte

Le matériel acoustique susceptible d'être utilisé pour la reconnaissance de locuteur ne se limite pas à un ensemble de mots de passe; il est au contraire très divers. Toutefois, il est possible de tenter d'en donner une classification en termes de contraintes pour la personne dont l'identité doit être reconnue:

- Mot de passe
  Texte fixe, prononcé de manière invariable. La machine connaît le texte qu'elle va entendre, et l'a déjà entendu plusieurs fois. La coopération du locuteur est requise.
- Texte imposé
  Texte variable, prononcé de manière neutre. La machine connaît le texte qu'elle va entendre, mais l'entend peut-être pour la première fois. La coopération du locuteur est facultative.
- Vocabulaire restreint
  Texte variable, prononcé de manière neutre. La machine connaît les mots du texte qu'elle va entendre, mais pas forcément leur ordre. La coopération du locuteur est facultative.
- Indépendance du texte
  Texte variable, prononcé de manière neutre. La machine suppose cependant qu'elle a affaire à une unique personne. La coopération du locuteur n'est plus requise.
- Indépendance totale
  Plusieurs locuteurs peuvent s'exprimer simultanément ou en séquences (même très courtes); un aspect émotionnel peut intervenir (cris, rires, pleurs, chants); d'autres sources sonores peuvent polluer le matériel acoustique (musique, bruits de circulation); l'environnement peut être responsable d'effets perturbateurs (échos); le milieu acoustique peut ne pas être conventionnel (plongeurs en atmosphère d'hélium).

Nous avons choisi de mener nos expériences de façon indépendante du texte. Nous avons utilisé à cet effet un matériel acoustique constitué d'une série de mots courts, identiques pour tous les locuteurs, et dont l'ordre était aléatoire. Cependant, nous n'avons pas fait usage des informations lexicales disponibles dans ce signal de parole. En particulier, à aucun moment le système de reconnaissance de locuteurs que nous proposons n'est conscient du fait que ce qu'il entend peut être découpé en mots distincts; *a fortiori,* il ne sait pas de quels mots il s'agit. Par conséquence, l'indépendance du texte est bien respectée.

### 1.3. Mode de reconnaissance

La tâche de reconnaissance elle-même peut être de l'une des trois natures suivantes:

- Vérification par acceptation
  Un locuteur déclare son identité avant de soumettre un texte oral susceptible de confirmer ou d'infirmer ses prétentions. Le texte soumis et uniquement la référence du locuteur en cause sont comparés. La décision finale est binaire: rejet ou acceptation.
- Vérification par rejet
  Un locuteur déclare son identité avant de soumettre un texte oral susceptible de confirmer ou d'infirmer ses prétentions. Le texte soumis est comparé avec toutes les références connues, qui doivent en principe unanimement déclarer le locuteur comme imposteur, à l'exclusion de celle dont l'identité a été revendiquée. La décision finale, combinaison des décisions partielles, est binaire: rejet ou acceptation.
- Identification 1 à $n$
  L'identité d'un locuteur est établie sur la base d'une comparaison de ses caractéristiques orales avec celles de tous les locuteurs connus de la machine. Le meilleur appariement désigne l'identité du locuteur. La décision finale est n-aire.
- Identification 1 à $(n+1)$
  Le fonctionnement est identique à celui d'une identiffication 1 à $n$, à ceci près que si aucun appariement d'une qualité suffisante n'est trouvé, alors le locuteur est déclaré inconnu de la machine. La décision finale est $(n+1)$-aire.

Nous avons choisi de mener nos expériences dans un contexte de vérification du locuteur par acceptation.

### 1.4 Figure de mérite

L'évaluation des performances de la méthode qu'utilise la machine pour reconnaître un locuteur peut se faire sur la base de plusieurs critères. Tout d'abord, il est important de préciser les conditions d'acquisition du matériel acoustique; en règle générale, les meilleurs résultats sont obtenus

avec l'utilisation d'un mot de passe, les pires s'observent lorsque les locuteurs ne respectent aucune contrainte, quelle qu'elle soit. Ensuite, la nature de la tâche de reconnaissance influe sur le résultat; ainsi, la tâche d'identification 1 à $n$ fournit souvent de meilleurs résultats (univers de locuteurs clos) que les tâches d'identification 1 à $(n+1)$ (univers de locuteurs ouvert) et de vérification par rejet ou par acceptation.
Enfin, la figure de mérite elle-même est variable. Elle est le plus souvent basée sur l'estimation *a posteriori* des probabilités d'erreur, qui couvrent deux catégories: un imitateur réussit son imposture (fausse acceptation), ou un utilisateur agréé est rejeté par le système (faux rejet). On distingue trois principales figures de mérite:

- Taux moyen minimal de faux rejets et de fausses acceptations (minimum average false reject and false acceptance rate, MAFRA)
  Ce critère [12] génère les meilleures figures de mérite (en pourcentage d'erreur) lorsque l'apparition d'un imposteur ou d'un locuteur licite est équiprobable.
- Taux d'erreur équitable (equal error rate, EER)
  Ce critère définit un taux d'erreur tel que la probabilité de se tromper, en déclarant comme imposteur un honnête homme, soit identique à la probabilité de se tromper, en déclarant comme honnête homme un imposteur.
- Taux constant de fausses acceptations (constant false acceptance rate, CFA)
  Ce critère, basé sur la mesure du taux de faux rejets pour un taux donné de fausses acceptations, génère usuellement les moins bonnes figures de mérite. Il est important de constater qu'il est néanmoins lié à la valeur d'un paramètre libre, et que deux CFA ne peuvent être comparés que si leur paramètre associé est identique.

Le critère que nous avons retenu est utilisé par la majorité des auteurs; il s'agit du taux d'erreur équitable.

## 2. Espace des vecteurs cepstraux

Une des descriptions du signal acoustique reconnue comme des plus efficaces pour la reconnaissance du locuteur est le cepstre à court-terme [6, 10, 18, 23, 24, 26]. Nous fonderons donc sur ce paramètre la série d'expériences décrite ici. Pour mémoire, le cepstre réel d'un signal est la transformée de Fourier inverse du logarithme naturel du module de la transformée de Fourier du signal.

$$c_f(k) = \frac{1}{N}\sum_{m=0}^{N-1} \ln\left|\sum_{n=0}^{N-1} x(n)\exp(-j\frac{2\pi}{N}nm)\right| \exp(j\frac{2\pi}{N}km) \quad k \in [0, N-1] \quad (1)$$

Le domaine de définition du cepstre est un axe temporel gradué en unités de quéfrence. C'est une description des périodes (inverses de fréquences) que l'on trouve dans le signal.

### 2.1 Calcul du cepstre

Deux particularités importantes rendent populaire la représentation cepstrale d'un signal de parole. D'une part, et par construction, elle se prête bien à la déconvolution homomorphique; d'autre part, outre la voie directe exposée en (1), existe une possibilité de calcul moins coûteuse en nombre d'opérations. Il s'agit de se livrer à une extraction des coefficients de prédiction linéaire $a(n)$ (Linear Prediction Coefficients, LPC) du signal, et de les transformer par récurrence:

$$c_p(n) = \begin{cases} 0 & n = 0 \\ -a(1) & n = 1 \\ -a(n) - \sum_{k=1}^{n-1} \frac{k\, c_p(k) a(n-k)}{n} & n > 1 \end{cases} \quad (2)$$

où $a(n)$ est le $n$-ième coefficient de prédiction linéaire (filtre inverse). Ces coefficients s'obtiennent par la solution du système ci-dessous de $P$ équations linéaires, où $P$ est l'ordre d'analyse:

$$a(n) = \begin{cases} 1 & n = 0 \\ \sum_{k=1}^{P} a(k) R(|n-k|) = R(j) & j \in [1, P] \\ 0 & j > P \end{cases} \quad (3)$$

La matrice $R$ des coefficients biaisés d'autocorrélation s'obtient par

$$R(k) = \sum_{n=0}^{N-1-k} h(n)h(n+k) \qquad k \in [0, P] \quad (4)$$

où les valeurs du signal de parole préaccentué $y(n)$, $n \in [0, N-1]$, ont été multipliées par une fenêtre de Bartlett.

$$h(n) = y(n)\left(1 - \frac{2|n - N/2|}{N}\right) \qquad n \in [0, N-1] \quad (5)$$

On observe que la pente spectrale moyenne d'un signal de parole est généralement négative. De sorte à accentuer la partie haute fréquence du spectre, le signal est traité par un filtre linéaire de transmittance $1 - \alpha \cdot z^{-1}$, appelé filtre de préaccentuation. En théorie, il faudrait adapter $\alpha$ au cours du temps, l'optimum étant atteint quand sa valeur est égale au coefficient du filtre inverse d'ordre 1. En pratique, la valeur de $\alpha$ n'est pas critique. Nous imposerons $\alpha = 0.95$ comme taux de préaccentuation du signal $s(n)$, $n \in [0, N-1]$.

$$y(n) = \begin{cases} 0 & n = 0 \\ s(n) - \alpha\, s(n-1) & n \in [1, N-1] \end{cases} \quad (6)$$

Avec cette méthode de calcul du cepstre par LPC, il est possible de ne retenir du signal que les caractéristiques importantes telles que par exemple les formants. En effet, un certain lissage spectral s'obtient automatiquement lorsque se fait le choix de l'ordre d'analyse $P$. Il s'ensuit que nous ne bénéficierons pas, dans notre cas, des avantages liés à la connaissance du fondamental $F_0$ d'un locuteur, même s'il est facile de se convaincre que cette caractéristique simple permet de distinguer, dans la plupart des cas, un homme d'une femme ou d'un enfant.

### 2.2 Extraction des vecteurs cepstraux

Nous avons décidé de procéder comme suit pour l'extraction des paramètres:

- Acquisition, $f_s = 8.0$ [KHz], $f_c = 3.4$ [KHz].
- Préaccentuation, $\alpha = 0.95$.
- Découpage du signal en tranches de 30 [ms] (240 échantillons), avec une partie de 20 [ms] commune à deux tranches voisines (recouvrement de 2/3).
- Multiplication de chaque tranche par une fenêtre de Bartlett.

- Analyse LPC 14 pôles par la méthode de Levinson.
- Transformation des vecteurs LPC à 14 composantes en vecteurs cepstraux à 14 composantes.

Les vecteurs cepstraux obtenus seront la base unique des méthodes que nous allons utiliser ici pour extraire des données l'identité du locuteur. Nous n'avons pas cherché à rendre plus efficace la génération de ces paramètres; ainsi par exemple, rien n'a été tenté pour éviter de traiter comme parole les plages de silence qui apparaissent dans le signal. Nous n'avons pas plus réalisé ou exploité une décision concernant le voisement de chaque tranche.

# 3. Méthodes

Nous avons testé trois méthodes génériques différentes permettant de quantifier l'appariement de deux locutions représentées par l'évolution temporelle de leur cepstre à court-terme. Une quatrième méthode a aussi été testée, qui découle directement de l'une des précédentes. Enfin, la mise en commun de ces quatre résultats permet de se rendre compte de l'efficacité de notre approche. Nous avons à disposition un ensemble $\mathbb{L}$ de $L$ locutions $\mathbb{L}_i$

$$\mathbb{L} = \{\mathbb{L}_i \,|\, i \in [0, L-1]\} \tag{7}$$

constituées chacune de $L_i$ vecteurs cepstraux $\boldsymbol{C}_{i,j}$ à $P$ composantes.

$$\mathbb{L}_i = \{\boldsymbol{C}_{i,j} \,|\, j \in [0, L_i - 1]\} \qquad i \in [0, L-1] \tag{8}$$

$$\boldsymbol{C}_{i,j} = \begin{pmatrix} c_{i,j,0} \\ \vdots \\ c_{i,j,P-1} \end{pmatrix} \qquad i \in [0, L-1] \qquad j \in [0, L_i - 1] \tag{9}$$

Dans notre cas, $L=80$, $L_i=1531$, $P=14$. De plus nous connaissons *a priori* l'appariement vrai entre une locution et un locuteur.

### 3.1 Cepstre moyen

De par les propriétés homomorphiques de l'analyse cepstrale, il est possible d'affirmer que la moyenne temporelle de tous les cepstres d'une locution représente, grossièrement, la mise en cascade d'abord du filtre créé par le conduit vocal au repos du locuteur, et ensuite de toutes les fonctions de transfert liées aux conditions d'acquisition (acoustique de la salle, réponse en fréquence du microphone, de l'enregistreur, du matériau magnétique, du filtre de garde, etc.). Ces transformations ayant été identiques pour tous les locuteurs, nous pouvons espérer que les différences entre les cepstres moyens issus de deux locutions décrivent directement la différence morphologique entre les deux locuteurs. La mesure scalaire de cette différence a été calculée par une métrique euclidienne.
Caractérisons une locution par son cepstre moyen.

$$\boldsymbol{L}_i = \frac{1}{L_i} \sum_{j=0}^{L_i - 1} \boldsymbol{C}_{i,j} \qquad i \in [0, L-1] \tag{10}$$

Quantifions la différence entre deux locutions par une métrique euclidienne.

$$d_1(\boldsymbol{L}_i, \boldsymbol{L}_k) = \|\boldsymbol{L}_k - \boldsymbol{L}_i\| \qquad i, k \in [0, L-1] \tag{11}$$

Décidons d'un seuil de vérification $\mu$ capable de nous dire s'il faut accepter ou rejeter l'affirmation selon laquelle les deux cepstres moyens $\boldsymbol{L}_i$ et $\boldsymbol{L}_k$ sont issus du même locuteur.

$$\text{décision} = \begin{cases} \text{acceptation} & \text{si } d_1(\boldsymbol{L}_i, \boldsymbol{L}_k) < \mu \\ \text{rejet} & \text{sinon} \end{cases} \tag{12}$$

### 3.2 Erreur accumulée de quantification vectorielle

La deuxième méthode testée examine plus en détail l'ensemble des cepstres à disposition [18, 23, 24]. L'idée de base est de se dire que, parmi eux, il en est certains qui décrivent un état stable (à court-terme) du signal vocal, tandis que d'autres, en moins grand nombre, représentent des états transitoires. Or, l'expérience montre que les états stables mentionnés sont en nombre réduit (de l'ordre de grandeur d'une ou de quelques centaines); en outre, chaque locuteur en possède un registre qui lui est propre [13]. Le but que nous nous fixons est tout d'abord d'identifier ce registre d'états stables, appelé vocabulaire, dont les mots sont les états stables. Cette partie de la phase d'apprentissage se nomme classification. Ensuite, en phase de reconnaissance, nous tenterons de reconstruire une locution complète à l'aide des seuls mots du vocabulaire; cette étape se nomme quantification vectorielle. La différence entre la locution originale et la locution reconstruite sera notre mesure de dissimilitude. En principe, le vocabulaire d'un locuteur devrait être efficace pour reconstruire toute locution émise par lui-même, et inefficace pour reconstruire une locution émise par un imposteur.

#### 3.2.1 Etablissement du vocabulaire

Nous allons découper une locution $\mathbb{L}_i$ en sous-ensembles regroupant des vecteurs cepstraux proches, puis représenter chaque sous-ensemble par un unique vecteur cepstral de substitution [14, 18, 24]. L'ensemble de ces représentants constituera le vocabulaire associé à la locution $\mathbb{L}_i$ (et par conséquent au locuteur qui l'a produite). Plusieurs méthodes de classification peuvent être utilisées pour réaliser ce but; elles se distinguent d'une part par les métriques de dissimilitude au sein d'un sous-ensemble ou entre deux sous-ensembles, d'autre part par la façon d'extraire un représentant d'un sous-ensemble, et enfin par des considérations plus globales telles que décision *a priori* ou *a posteriori* de la taille du vocabulaire, condition d'arrêt pour les méthodes qui font usage de techniques itératives de convergence, etc.
Nous avons choisi d'utiliser une technique itérative de classification, nommée H-means, avec le barycentre d'un sous-ensemble dans le rôle de représentant de ce sous-ensemble. Cette technique cherche à minimiser le rapport entre, d'une part, la somme des distances euclidiennes entre les éléments d'un sous-ensemble et son représentant, et, d'autre part, la somme des distances entre les représentants eux-mêmes. La convergence de cette méthode est assurée (même si le minimum atteint peut n'être qu'un minimum local). Le nombre de classes $Q$ doit être imposé *a priori;* dans notre cas, $Q=32$.
Soit un ensemble $\mathbb{L}_i$ de vecteurs cepstraux, et soit une partition $\mathbb{P}_j^t$ de cet ensemble, donnée par ses représentants, où $t$ désigne l'indice d'itération:

$$\mathbb{P}_i^t = \{\boldsymbol{P}_{i,p}^t | p \in [0, Q-1]\} \qquad i \in [0, L-1] \tag{13}$$

Comme condition initiale, nous extrairons $\mathbb{P}_i^0$ de $\mathbb{L}_i$ par une méthode quelconque.

$$\mathbb{P}_i^0 \subset \mathbb{L}_i \tag{14}$$

De façon itérative, nous assignerons chaque élément $\boldsymbol{L}_{i,j}$ au plus proche représentant $\boldsymbol{P}_{i,q}^t$ de la partition $\mathbb{P}_i^t$, de sorte à construire une nouvelle répartition $\mathbb{P}_{i,q}^{t+1}$ des éléments

$$q = \underset{p=0}{\overset{Q-1}{\operatorname{argmin}}}\, d_1(\boldsymbol{L}_{i,j}, \boldsymbol{P}_{i,q}^t) \quad \Rightarrow \quad \boldsymbol{L}_{i,j} \in \mathbb{P}_{i,q}^{t+1} \tag{15}$$

$$i \in [0, L-1] \qquad j \in [0, L_i - 1]$$

puis nous établirons les nouveaux représentants $\boldsymbol{P}_{i,p}^{t+1}$ par le calcul du barycentre de chaque répartition $\mathbb{P}_{i,p}^{t+1}$

$$\boldsymbol{P}_{i,p}^{t+1} = \frac{\sum_{m=0}^{\mathrm{card}(\mathbb{P}_{i,p}^{t+1})-1} \mathrm{elem}(m, \mathbb{P}_{i,p}^{t+1})}{\mathrm{card}(\mathbb{P}_{i,p}^{t+1})} \tag{16}$$

$$i \in [0, L-1] \qquad p \in [0, Q-1]$$

où la notation $\mathrm{elem}(n, \mathbb{E})$ désigne le $n$-ième élément de l'ensemble $\mathbb{E}$. Si la partition de $\mathbb{L}_i$ ne s'est pas modifiée, alors nous prétendons que la solution a été trouvée; sinon, nous passons à l'itération suivante.

$$\text{itération} = \begin{cases} \text{arrêt} & \text{si } \boldsymbol{P}_{i,p}^t = \boldsymbol{P}_{i,p}^{t+1} \\ t \rightarrow t+1 & \text{sinon} \end{cases} \qquad \forall p \in [0, Q-1] \tag{17}$$

### 3.2.2 Stratégie de classification

Nous savons que la méthode de classification H-means converge toujours; cependant, cette convergence n'est pas absolue: il se peut que le minimum atteint ne soit que local. Ce fait nous contraint à choisir une bonne partition initiale $\mathbb{P}^0$. Nous y parvenons en trois étapes.

La première étape fournit une contribution indépendante pour chaque locution $\mathbb{L}_i$. Nous commençons par établir une partition initiale $\mathbb{P}_i^0$ simplement en considérant les $Q$ premiers vecteurs de la locution concernée, puis nous calculons une partition finale $\mathbb{P}_i$.

La deuxième étape établit une contribution globale pour toutes les locutions. Nous quantifions vectoriellement chaque locution $\mathbb{L}_i$ par le vocabulaire $\mathbb{P}_i$ obtenu lors de la première étape, en ne conservant qu'une fraction des données (par exemple 10%), mais en amalgamant les résultats de sorte à créer une grande locution finale $\mathbb{X}$. Cette réduction du volume de données (de $L \cdot L_i$ à $10\% \cdot (L \cdot L_i)$) est souhaitable pour deux raisons: d'une part, la tâche de classification s'en trouve accélérée d'autant, et, d'autre part, certains mots rares du vocabulaire disparaissent, laissant la place à une représentation plus fine des mots les plus fréquents.

La troisième étape consiste à mener le travail de classification sur l'amalgame $\mathbb{X}$, en utilisant comme partition initiale ses premiers vecteurs. La partition $\mathbb{P}$, obtenue après convergence, sert enfin de partition initiale unique $\mathbb{P}^0$, identique pour chaque locution $\mathbb{L}_i$, car nous admettons que cette partition $\mathbb{P}^0$ est de bonne qualité, propre à générer pour chaque locution $\mathbb{L}_i$ une partition finale $\mathbb{P}_i$ proche du minimum absolu.

### 3.2.3 Mesure de dissimilitude

Nous supposons connaître le vocabulaire d'un locuteur donné, et nous allons tenter d'exprimer une locution avec les mots de ce vocabulaire. A l'issue de cette traduction, la locution aura été modifiée; notre but est d'établir une mesure quantitative de cette modification. La métrique que nous avons retenue consiste à calculer la somme normalisée des distances euclidiennes minimales entre les vecteurs cepstraux de la locution et ceux du vocabulaire.

Soit une locution $\mathbb{L}_i$ et un vocabulaire (partition finale) $\mathbb{P}_k$. L'erreur accumulée de quantification vectorielle est

$$d_2(\mathbb{L}_i, \mathbb{P}_k) = \frac{1}{L_i} \sum_{j=0}^{L_i-1} \min_{p=0}^{Q-1} d_1(\boldsymbol{L}_{i,j}, \boldsymbol{P}_{k,p}) \qquad i,k \in [0, L-1] \tag{18}$$

### 3.3. Cepstre différentiel

La méthode du cepstre différentiel est en fait exactement la même que la précédente; seules les caractéristiques utilisées changent. Nous avons en effet remplacé les vecteurs cepstraux par des vecteurs estimant leur dérivée temporelle (à ne pas confondre avec la dérivée quéfrentielle). Ainsi, nous profitons pleinement des possibilités de la déconvolution homomorphique, puisque nous supprimons automatiquement les contributions de tout état stable, même à court-terme. Le signal qui résulte de la dérivation temporelle des vecteurs cepstraux est donc une représentation directe des états transitoires. Etant donné que nous avons jusqu'à maintenant surtout profité des plages de stabilité du signal, nous espérons que l'utilisation de ces nouveaux paramètres fournira des résultats très décorrélés avec les autres [7].

Soit la locution $\mathbb{L}_i$ de l'équation (8) et ses vecteurs cepstraux (9). Nous estimerons les vecteurs cepstraux différentiels par:

$$\dot{\boldsymbol{C}}_{i,j} = \boldsymbol{C}_{i,j+1} - \boldsymbol{C}_{i,j-1} \qquad i \in [0, L-1] \qquad j \in [1, L_i - 2] \tag{19}$$

Ceci nous fournit un vecteur cepstral différentiel chaque 10 [ms], dont la valeur a été estimée sur une plage de 30 [ms]. L'ensemble de ces vecteurs $\dot{\boldsymbol{C}}_{i,j}$ ira se substituer aux $\boldsymbol{C}_{i,j}$ déjà utilisés dans les deux méthodes connues; il reste encore à décider laquelle nous allons employer. La première se prête mal à une telle substitution, car la moyenne des $\dot{\boldsymbol{C}}_{i,j}$ est presque identiquement nulle, chaque terme de la somme annulant le terme précédent, aux bornes près. La seconde est plus propice; nous allons donc l'utiliser, tout en conservant identiques les paramètres, même s'il est clair que maintenir la taille du vocabulaire n'est pas sans risques. La raison en est que s'il existe $Q$ états stables, alors il existe $Q \cdot (Q-1)$ transitions possibles entre ces états, c'est-à-dire bien plus (dès que $Q > 1$). Nous n'avons pas tenu compte de ce fait, et nous avons conservé $Q = 32$.

Soit une locution de test $\dot{\mathbb{L}}_i$ et un vocabulaire de référence (partition finale) $\dot{\mathbb{P}}_k$. L'erreur accumulée de quantification vectorielle pour les cepstres différentiels est:

$$d_3(\dot{\mathbb{L}}_i, \dot{\mathbb{P}}_k) = \frac{1}{L_i} \sum_{j=0}^{L_i-1} \min_{p=0}^{Q-1} d_1(\dot{\boldsymbol{L}}_{i,j}, \dot{\boldsymbol{P}}_{k,p}) \qquad i,k \in [0, L-1] \tag{20}$$

### 3.4 Histogrammes

Nous en arrivons à la dernière méthode indépendante, que nous appellerons méthode des histogrammes. Elle repose sur l'espoir qu'un locuteur se distingue de ses pairs par certains tics de langage, par exemple par l'utilisation plus fréquente de certains cepstres, à l'exclusion de certains autres; bien qu'elle se rapproche de [22] dans son principe, c'est une méthode inédite.

De la technique d'accumulation d'erreur de quantification vectorielle nous retiendrons la phase de classification, que nous appliquerons à l'ensemble de tous les vecteurs cepstraux de tous les locuteurs disponibles. Le vocabulaire que nous en obtiendrons sera appelé vocabulaire global, par contraste avec les vocabulaires personnels précédemment établis. Ce vocabulaire a la prétention de représenter non plus un seul locuteur, mais l'humanité toute entière; il s'ensuit que le nombre de classes est obligatoirement plus élevé. Dans notre cas, nous avons posé $Q = 256$.
Ensuite, nous construirons une fonction de répartition qui associe, pour un locuteur donné, une probabilité d'apparition à tout mot du vocabulaire global; ce sera la référence du locuteur. En phase de reconnaissance, une locution inconnue sera quantifiée vectoriellement par le vocabulaire global, et le résultat de cette quantification sera analysé de sorte à établir sa fonction de répartition. Finalement, la distance entre la fonction de répartition obtenue et celle de son auteur potentiel sera mesurée; un seuil $\mu$, choisi judicieusement lors de la phase d'apprentissage, permettra de décider si l'on a affaire à un imposteur ou non.
Soit une locution $\mathbb{L}_i$ et le vocabulaire global $\mathbb{P}$. La fonction de répartition associée est

$$\boldsymbol{H}_i = \begin{pmatrix} \frac{\mathrm{card}(\mathbb{L}_i^q \cap \mathbb{P}_0)}{L_i} \\ \vdots \\ \frac{\mathrm{card}(\mathbb{L}_i^q \cap \mathbb{P}_{Q-1})}{L_i} \end{pmatrix} \qquad i \in [0, L-1] \tag{21}$$

où $\mathbb{L}_i^q$ est la version de $\mathbb{L}_i$ quantifiée vectoriellement, et où $\mathrm{card}(\mathbb{L}_i^q \cap \mathbb{P}_p)$ représente le nombre d'éléments $\boldsymbol{P}_p$ trouvés dans la locution quantifiée $\mathbb{L}_i^q$.

$$\mathbb{L}_i^q = \{ \boldsymbol{P}_{p_j} | (p_j = \underset{p=0}{\overset{Q-1}{\operatorname{argmin}}} d_1(\boldsymbol{L}_{i,j}, \boldsymbol{P}_p)) \wedge (j \in [0, L_i - 1]) \} \; i \in [0, L-1] \tag{22}$$

La distance euclidienne entre deux locutions est

$$d_4(\boldsymbol{H}_i, \boldsymbol{H}_k) = \| \boldsymbol{H}_k - \boldsymbol{H}_i \| \qquad i, k \in [0, L-1] \tag{23}$$

## 4. Discriminant linéaire de Fisher

Nous allons maintenant mettre ensemble tous les résultats obtenus, en les combinant de sorte à en extraire les meilleures performances possibles; le point qu'il reste à résoudre est la nature de cette combinaison [2, 15, 16, 20]. Nous avons décidé de conserver les méthodes précédentes intactes; seules les distances qu'elles calculent seront combinées. Nous voulons donc séparer, au moyen d'un hyperplan, l'espace quadridimensionnel des distances en deux hyper-espaces qui déterminent les domaines intra-locuteur et inter-locuteur. Cette façon de procéder a l'avantage de la simplicité, car le test d'appartenance à l'un des deux domaines se calcule par un simple produit scalaire de deux vecteurs; un autre avantage est l'existence d'une technique éprouvée, déterministe, qui permet de traiter ce cas.
Cette méthode se nomme discriminant linéaire de Fisher [17]; elle a pour objectif la maximisation du critère donné par le rapport entre la distance inter-classe et la somme des dispersions intra-classe. Pour calculer ces caractéristiques, nous projetons tout d'abord chaque point sur une droite arbitraire passant par l'origine, et nous repérons ensuite la position du point par un scalaire qui exprime la distance entre sa projection et l'origine. C'est par rapport à cette dernière grandeur que nous définissons la dispersion intra-classe et inter-classe. La maximisation du critère se fait en terme d'orientation de la droite de projection; l'hyper-plan mentionné y est perpendiculaire, sis à une distance de l'origine qui est directement le seuil de décision de classification entre domaine intra-locuteur et domaine inter-locuteur.
Soit $\mathbb{D}_\cup$ l'ensemble des dissimilitudes intra-locuteur obtenues par les quatre méthodes précédentes, et $\mathbb{D}_\cap$ l'ensemble des dissimilitudes inter-locuteur

$$\mathbb{D}_\cup = \{ \boldsymbol{d}_j | j \in [0, D_\cup - 1] \} \tag{24}$$

$$\mathbb{D}_\cap = \{ \boldsymbol{d}_k | k \in [0, D_\cap - 1] \} \tag{25}$$

où $D_\cup$, $D_\cap$ sont le nombre d'expériences menées dans chaque cas, et $\boldsymbol{d}_i$ les vecteurs quadridimensionnels obtenus.

$$\boldsymbol{d}_i = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{pmatrix} \qquad i \in [0, D_{\{\cap,\cup\}} - 1] \tag{26}$$

Décrivons par un vecteur directeur $\boldsymbol{w}$ une droite passant par l'origine. Chaque vecteur de dissimilitude se projette sur cette droite, à une distance $y$ de l'origine.

$$y_i = \boldsymbol{w}^T \boldsymbol{d}_i \qquad i \in [0, D_{\{\cap,\cup\}} - 1] \tag{27}$$

Nous estimerons la différence inter-classe $\delta_\cap$ par:

$$\delta_\cap = \boldsymbol{w}^T S_B \boldsymbol{w} \tag{28}$$

$$S_B = (\boldsymbol{m}_\cup - \boldsymbol{m}_\cap)(\boldsymbol{m}_\cup - \boldsymbol{m}_\cap)^T \tag{29}$$

$$\boldsymbol{m}_\cup = \frac{1}{D_\cup} \sum_{j=0}^{D_\cup - 1} \boldsymbol{d}_j \tag{30}$$

$$\boldsymbol{m}_\cap = \frac{1}{D_\cap} \sum_{k=0}^{D_\cap - 1} \boldsymbol{d}_k \tag{31}$$

Nous estimerons la somme des dispersions intra-classe par $\delta_\cup$:

$$\delta_\cup = \boldsymbol{w}^T S_W \boldsymbol{w} \tag{32}$$

$$S_W = 0.5\, S_\cup + 0.5\, S_\cap \tag{33}$$

$$S_\cup = \frac{1}{D_\cup} \sum_{j=0}^{D_\cup - 1} (\boldsymbol{d}_j - \boldsymbol{m}_\cup)(\boldsymbol{d}_j - \boldsymbol{m}_\cup)^T \tag{34}$$

$$S_\cap = \frac{1}{D_\cap} \sum_{k=0}^{D_\cap - 1} (\boldsymbol{d}_k - \boldsymbol{m}_\cap)(\boldsymbol{d}_k - \boldsymbol{m}_\cap)^T \tag{35}$$

Le critère à maximiser est:

$$J(\boldsymbol{w}) = \frac{\delta_\cap}{\delta_\cup} \tag{36}$$

La solution est:

$$\boldsymbol{w} = S_W^{-1}(\boldsymbol{m}_\cup - \boldsymbol{m}_\cap) \tag{37}$$

Nous avons obtenu le discriminant linéaire de Fisher, qui maximise le rapport entre l'écart inter-classe et la somme

des dispersions intra-classe lorsque l'on projette sur une droite les points d'un espace multidimensionnel. Avant de passer plus loin, il paraît nécessaire de formuler quelques commentaires.
Tout d'abord, constatons que l'espace dans lequel nous travaillons n'est pas homogène. Afin d'y remédier, nous avons décidé de normaliser chaque composante du vecteur $\boldsymbol{d}$ par un facteur de pondération égal à l'inverse du seuil de dissimilitude $\mu_{EER}$ établi par la méthode correspondante. Ceci tend à harmoniser numériquement les contributions de chaque dimension. Pour être complet, il aurait encore été nécessaire de s'intéresser aux moments d'ordre deux (égalisation des variances) et supérieurs, mais l'apport en a été jugé insuffisant, eu égard à la complexité supplémentaire.
Ensuite, nous nous souviendrons que les définitions (34) et (35) des matrices de dispersion inter-classe font appel aux coefficients de pondération $1/D_{\cup}$ et $1/D_{\cap}$. Faire conjointement usage des coefficients 0.5 à l'expression (33) signifie que l'on suppose *a priori* que la probabilité d'apparition d'un imposteur et celle d'un honnête homme sont égales entre elles. Si l'on désire au contraire estimer ces probabilités *a posteriori,* alors donner une valeur unité à tous ces coefficients de pondération est la voie la plus simple, que l'on trouve parfois dans la littérature [5].
Enfin, il est utile de mentionner le fait que les distances mesurées entre un point projeté sur la droite et l'origine peuvent devenir parfois négatives. Ce fait est susceptible de poser des problèmes à toute méthode qui présuppose qu'une distance est un scalaire positif ou nul, condition qui n'est pas respectée par le discriminant linéaire de Fisher.

## 5. Expérimentation

La discussion du paragraphe 1.2 nous a rappelé les conditions nécessaires à la reconnaissance du locuteur qu'impose le choix du matériel vocal. Formellement, parmi les différents modes de reconnaissance, nous avons adopté la contrainte du vocabulaire restreint pour mener nos expériences, mais sans jamais tenter de bénéficier des avantages liés à cette réduction de l'univers acoustique. Ainsi, c'est le principe d'indépendance du texte qui est respecté. Nous avons choisi de mener à bien la tâche de vérification par acceptation. Le critère d'évaluation retenu est le taux d'erreur équitable.

### 5.1 Matériel acoustique

Nous avons testé les algorithmes sur un univers de dix locuteurs, neuf hommes et une femme, qui couvrent la tranche d'âge comprise entre vingt et quarante ans. La procédure d'acquisition du signal acoustique a demandé à chaque locuteur de lire une série de vingt mots courts, tous différents, dont l'ordre était aléatoire au sein de chaque série; la liste de ces mots est donnée à la figure 1. Chacun des dix locuteurs a produit, en une session unique, huit locutions; la durée de chaque locution était de quinze secondes. La reproductibilité de cette durée, quel que soit le locuteur, a été obtenue par un mécanisme de dictée: sur un écran s'affichaient, avec un rythme imposé, les mots à prononcer. Le locuteur avait cependant toute liberté quant à l'instant de départ d'une série.

$S = 10$ *Nombre de locuteurs* (38)
$L = 80$ *Nombre de locutions* (39)

| | | | |
|---|---|---|---|
| Un | Deux | Trois | Quatre |
| Cinq | Six | Sept | Huit |
| Neuf | Dix | Onze | Douze |
| Treize | Quinze | Seize | Vingt |
| Trente | Cent | Mille | Chiffre |

**Figure 1: Liste des mots contenus dans chaque locution.**

L'enregistrement des locutions s'est fait dans une salle calme, sans précautions acoustiques particulières. La distance entre le locuteur et le microphone Superscope EC-7 utilisé était libre de toute contrainte; elle était en général de cinquante centimètres environ. Le signal acoustique était transmis à un enregistreur Marantz SD275, muni d'une cassette vierge Maxell UDI46, avec utilisation d'un réducteur de bruit Dolby c.
La numérisation de ce signal a été réalisée par une carte DataTranslation DT2821 couplée à un micro-ordinateur Teleprint TDC (compatible AT). Le filtre de garde était un Krohn-Hite 3343, ajusté en passe-bas RC de 3.4 [KHz] de fréquence de coupure, avec une atténuation asymptotique de 48 [dB/oct]; la cadence d'échantillonnage était de 8.0 [KHz], avec une résolution numérique linéaire de 12 [bit]. Ces valeurs ont été choisies telles qu'elles soient compatibles avec un signal transmis par canal téléphonique [8, 9, 11]; toutefois, aucune bande inférieure de fréquence n'a été coupée. Le gain du système de conversion analogique-numérique a été ajusté manuellement de cas en cas afin d'assurer que le signal n'était pas numériquement saturé, tout en couvrant une plage importante de la résolution (environ 11 [bit]).

### 5.2 Méthodologie d'établissement du seuil

L'établissement du seuil $\mu$ déterminant l'appartenance à un domaine intra-locuteur ou inter-locuteur d'une distance mesurée $d$ est une étape expérimentale importante. Deux écoles s'affrontent à ce sujet: la première veut que les seuils soient choisis *a priori;* ils pourront être adaptés par la suite, si nécessaire. Cette situation se rencontre presque toujours en phase d'exploitation, car il est alors exclu d'attendre que la quantité de données nécessaire au choix du seuil optimum soit réunie. Un autre avantage est que la connaissance objective des cas d'erreur n'est pas requise. La seconde école, très pratiquée en phase expérimentale, accepte de déterminer les seuils *a posteriori;* nous avons choisi cette voie. Nous devons en considérer trois variantes principales:

- Seuil intrinsèque à la méthode utilisée
  Nous choisissons un seuil unique, tel qu'il minimise le taux moyen d'erreur équitable sur toutes les expériences menées, pour une méthode donnée.
- Seuil intrinsèque au locuteur
  Nous offrons à chaque locuteur son propre seuil, tel qu'il minimise le taux moyen d'erreur équitable sur les expériences menées avec ses références personnelles, pour une méthode donnée (un locuteur peut posséder plusieurs références).
- Seuil intrinsèque à une référence
  Nous associons à chaque référence un seuil individuel, tel qu'il minimise le taux moyen d'erreur équitable sur les expériences menées avec cette référence seulement, pour une méthode donnée.

Le nombre de tests disponibles diminue de la première à la dernière variante, et la fiabilité des résultats diminue parallèlement. Cependant, les performances s'améliorent avec l'adaptation du seuil aux conditions locales. Ainsi, il est des références qui produisent naturellement de grandes distances, tout en étant ni plus ni moins discriminatoires que d'autres références produisant de petites distances. Nous présenterons donc nos résultats en choisissant des seuils intrinsèques aux références.

## 6. Résultats

Nous présenterons nos résultats sous deux formes: tout d'abord une visualisation graphique des distances, normalisées par la soustraction du seuil d'erreur équitable, intrinsèque à la référence, puis un tableau résumant les résultats par locuteur (les seuils ne dépendent toutefois toujours que des références). Dans ces tableaux, communément appelés matrices de confusion, nous associerons l'abscisse aux références et l'ordonnée aux locuteurs testés.
Dans les graphiques, nous regrouperons en abscisse les références $i \in [1,80]$ par paquet de huit, chaque paquet correspondant au locuteur $j$ des matrices de confusion, avec $j = (i-1)$ DIV 8. Les points de petite taille indiqueront les expériences du domaine inter-locuteur, et les points de taille plus grande celles du domaine intra-locuteur. Lorsque l'échelle le permettra, nous indiquerons la valeur correspondant à une distance nulle par un marqueur horizontal. La valeur du seuil, de par la normalisation, sera toujours d'ordonnée nulle.

### 6.1. Cepstre moyen

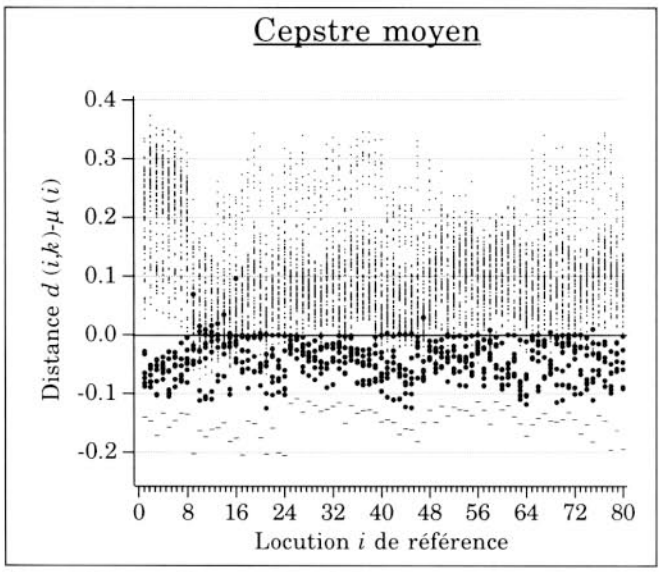


**Figure 2: Distances obtenues par la technique du cepstre moyen.**

X
Y Z = fa_quantified (Y by X)

| Y\X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 1 | 0 | 0 | 1 | 2 | 3 | 24 | 9 | 3 | 0 | 0 | 42 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 29 | 26 | 0 | 5 | 3 | 0 | 0 | 0 | 14 | 77 |
| 4 | 0 | 19 | 5 | 2 | 0 | 3 | 0 | 0 | 14 | 0 | 43 |
| 5 | 0 | 39 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 42 |
| 6 | 0 | 31 | 1 | 0 | 0 | 3 | 0 | 12 | 0 | 0 | 47 |
| 7 | 0 | 8 | 8 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 20 |
| 8 | 0 | 6 | 0 | 0 | 6 | 9 | 0 | 0 | 0 | 0 | 21 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| sum | 0 | 132 | 41 | 4 | 14 | 42 | 13 | 17 | 17 | 14 | 294 |

**Figure 3: Matrice de confusion pour le cepstre moyen, répartition des fausses acceptations pour 5760 tests.**

| ref | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| sum | 0 | 13 | 3 | 0 | 0 | 5 | 0 | 1 | 1 | 1 | 24 |

**Figure 4: Matrice de confusion pour le cepstre moyen, répartition des faux rejets pour 560 tests.**

Comparons notre résultat EER valant 4,7% à celui de [19], valant 5,9%. Nous en concluons que cette méthode fonctionne à satisfaction. Mieux encore, nous avons obtenu ce résultat sans suppression des plages de silence, sans orthogonalisation des paramètres cepstraux, et avec une durée d'apprentissage six fois plus courte que n'utilise la référence proposée.

### 6.2. Erreur accumulée de quantification vectorielle

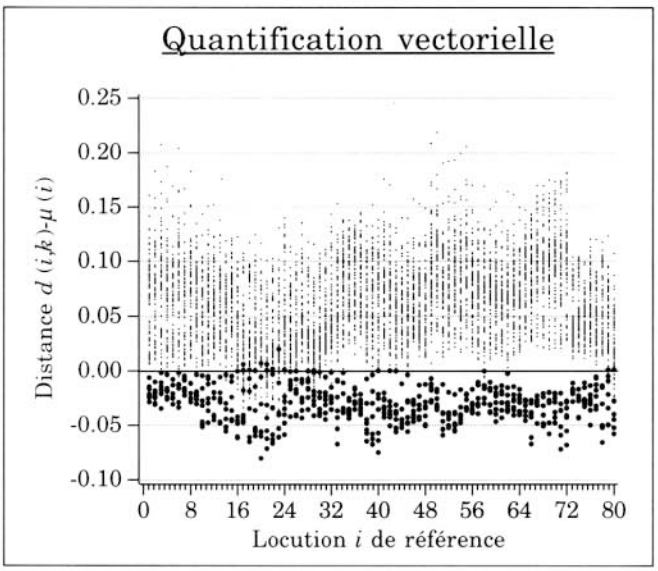


**Figure 5: Distances obtenues par la technique d'erreur accumulée de quantification vectorielle.**

X
Y Z = fa_quantified (Y by X)

| Y\X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 11 | 3 | 2 | 0 | 0 | 0 | 0 | 1 | 17 |
| 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 4 | 0 | 3 | 44 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 67 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 3 | 4 | 0 | 0 | 0 | 0 | 3 | 0 | 15 | 25 |
| 7 | 1 | 1 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 16 |
| 8 | 0 | 0 | 19 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 21 |
| 9 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| sum | 1 | 7 | 94 | 27 | 2 | 0 | 0 | 3 | 0 | 18 | 152 |

**Figure 6: Matrice de confusion pour la quantification vectorielle, répartition des fausses acceptations pour 5760 tests.**

| ref | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| sum | 0 | 0 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 13 |

**Figure 7: Matrice de confusion pour la quantification vectorielle, répartition des faux rejets pour 560 tests.**

Comparons notre résultat EER valant 2,5% (cepstre instantané) à celui de [18], valant 3,0%. Là encore nous pou-

vons nous montrer satisfaits, d'autant plus que la taille de notre vocabulaire est réduite à la moitié de celle de la référence citée, que nous n'avons utilisé que le tiers de la durée prise par [18] pour la tâche de classification, que notre métrique euclidienne n'offre pas les avantages de la métrique de Mahalanobis, qui tient compte des covariances de chaque paire de dimensions cepstrales, et enfin que le résultat de 2,5% annoncé concerne exclusivement le cepstre instantané, alors que le résultat de 3% de la référence [18] inclut déjà la combinaison du cepstre instantané et du cepstre différentiel.

### 6.3 Erreur accumulée de quantification vectorielle différentielle

```
    X
Y   Z = fa_quantified (Y by X)

 Y\X |    0    1    2    3    4    5    6    7    8    9 |  sum
-- = | -- = -- = -- = -- = -- = -- = -- = -- = -- = -- = | -- =
   0 |    0   39   15   50    7   64   25   57   10   13 |  280
   1 |    0    0    0    8    0   64    1   27    0    0 |  100
   2 |   43   48    0   48   37   64   48   56   34   46 |  424
   3 |    0    2    0    0    0   64    0    7    0    0 |   73
   4 |   35   59   42   59    0   64   41   62   39   37 |  438
   5 |    0    0    0    0    0    0    0    0    0    0 |    0
   6 |   19   33   17   35    3   64    0   61    2   14 |  248
   7 |    0    0    0    0    0   60    0    0    0    0 |   60
   8 |   40   53   38   56   40   64   42   61    0   41 |  435
   9 |   33   55   36   63   18   64   42   64   20    0 |  395
-- = | -- = -- = -- = -- = -- = -- = -- = -- = -- = -- = | -- =
 sum |  170  289  148  319  105  572  199  395  105  151 | 2453
```

**Figure 8: Matrice de confusion pour la quantification vectorielle différentielle, répartition des fausses acceptations pour 5760 tests.**

```
 ref |    0    1    2    3    4    5    6    7    8    9 |  sum
-- = | -- = -- = -- = -- = -- = -- = -- = -- = -- = -- = | -- =
 sum |   17   28   14   31   11   56   19   39   10   14 |  239
```

**Figure 9: Matrice de confusion pour la quantification vectorielle différentielle, répartition des faux rejets pour 560 tests.**

Notre résultat EER, valant 42,6%, peut être considéré comme mauvais. Cette piètre performance est même parfois catastrophique, comme par exemple pour le locuteur 5, où 572 tests sur 576 se sont révélés infructueux (domaine inter-locuteur), et où la totalité des 56 tests intra-locuteurs a conduit à l'échec! Par conséquent, nous n'avons pas inclus cette méthode dans la combinaison de Fisher. Formellement, nous récrirons l'expression (26) sans faire usage de la distance $d_3$.

La raison de la médiocrité de ces résultats est mal comprise. Une explication possible est que l'estimation (19) de la dérivée temporelle des vecteurs cepstraux est trop bruitée. Une autre thèse incrimine l'usage de la technique de classification H-means, qui tend à couvrir au mieux la totalité de l'espace vectoriel, mais sans tenir compte de la densité des vecteurs représentés par chaque mot du vocabulaire. Or, l'espace vectoriel des cepstres différentiels est loin d'être homogène, puisque les nombreuses plages stables du signal de parole s'accumulent près de l'origine. Il s'ensuit que la technique de classification choisie est peut-être inadaptée.

### 6.4 Histogrammes

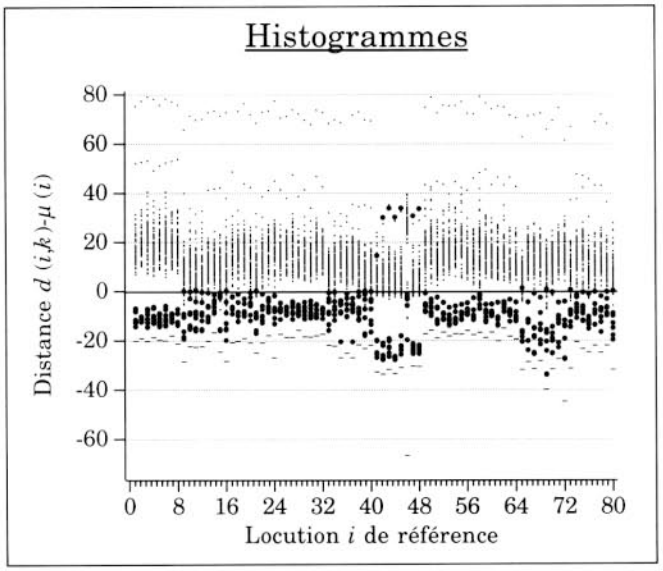


**Figure 10: Distances obtenues par la technique des histogrammes.**

```
    X
Y   Z = fa_quantified (Y by X)

 Y\X |    0    1    2    3    4    5    6    7    8    9 |  sum
-- = | -- = -- = -- = -- = -- = -- = -- = -- = -- = -- = | -- =
   0 |    0    0    0    0    0    0    0    0    0    0 |    0
   1 |    0    0    0    0   12   10    0    0    6    0 |   28
   2 |    0    0    0    0    7    1    0    0    1    1 |   10
   3 |    0    0    3    0    0   17    0    0    3   13 |   36
   4 |    0   13    2    0    0   10    0    0    8    2 |   35
   5 |    0    0    0    0    0    0    0    0    0    0 |    0
   6 |    0    8    0    0    0   16    0    0    1    5 |   30
   7 |    0   16    1    0    1   16    0    0    3    1 |   38
   8 |    0    0    0    0    0    0    0    0    0    0 |    0
   9 |    0    0    0    0    0    2    0    0    0    0 |    2
-- = | -- = -- = -- = -- = -- = -- = -- = -- = -- = -- = | -- =
 sum |    0   37    6    0   20   72    0    0   22   22 |  179
```

**Figure 11: Matrice de confusion pour la technique des histogrammes, répartition des fausses acceptations pour 5760 tests.**

```
 ref |    0    1    2    3    4    5    6    7    8    9 |  sum
-- = | -- = -- = -- = -- = -- = -- = -- = -- = -- = -- = | -- =
 sum |    0    3    1    0    1    7    0    0    2    2 |   16
```

**Figure 12: Matrice de confusion pour la technique des histogrammes, répartition des faux rejets pour 560 tests.**

La méthode de l'histogramme des vecteurs choisis par quantification vectorielle dans un vocabulaire universel est, semble-t-il, inédite, même s'il existe des méthodes apparentées [13, 22]. Notre résultat EER valant 3,0% ne leur est cependant pas comparable directement, ne serait-ce que parce que les figures de mérite utilisées [13], ainsi que les modes de reconnaissance choisis [22], sont différents. Cependant, nous ne pouvons qu'être satisfaits des résultats obtenus.

### 6.5 Discriminant linéaire de Fisher

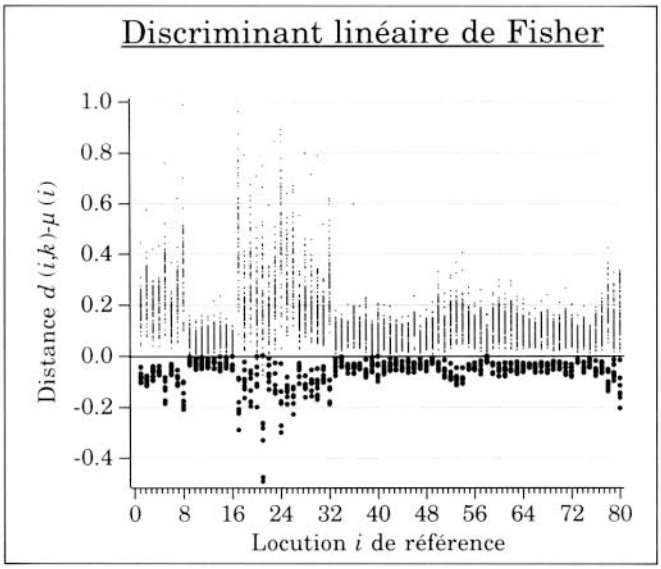


**Figure 13: Distances obtenues par la technique du discriminant linéaire de Fisher.**

X
Y Z = fa_quantified (Y by X)

| Y\X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 |
| 7 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| sum | 0 | 2 | 9 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 12 |

**Figure 14: Matrice de confusion pour le discriminant linéaire de Fisher, répartition des fausses acceptations pour 5760 tests.**

| ref | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| sum | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**Figure 15: Matrice de confusion pour le discriminant linéaire de Fisher, répartition des faux rejets pour 560 tests.**

L'analyse de ces résultats nous livre un taux moyen global d'erreur équitable valant 0,2% pour la technique de combinaison de trois méthodes par discriminant linéaire de Fisher (cepstre moyen, accumulation d'erreur de quantification vectorielle des cepstres instantanés, histogrammes). Nous pouvons comparer ce résultat avec celui de [2], valant 1,9%, où la combinaison de nombreuses méthodes a aussi été utilisée.

Nous ne nous autorisons cependant à montrer qu'un enthousiasme modéré devant nos bons résultats, car les conditions expérimentales nous sont particulièrement favorables:

- Session unique
  Nous n'avons tenu aucun compte de la variabilité des locuteurs au cours du temps, car les références, les données servant à l'établissement des seuils, ainsi que les locutions de test proviennent toutes de la même session. La séparation de ces données en lots à usage unique (apprentissage, seuil, test) est une opération qui ferait croître le taux d'erreur; l'introduction de sessions supplémentaires le ferait croître encore plus.
- Univers restreint
  Nous n'avons testé que dix locuteurs. La petite taille de cet ensemble rend peu probable l'apparition de deux voix difficilement séparables, ce qui nous facilite la tâche.
- Nombre de tests limité
  Chaque locuteur n'a produit que huit locutions, ce qui ne nous permet de mener que sept tests intra-locuteurs par locution (nous excluons le test entre une référence et la locution même qui l'a générée). Ainsi, l'analyse de Fisher consiste à séparer par un plan, dans un espace tridimensionnel, un nuage de sept points d'un nuage de soixante-douze points. Or, la probabilité d'un succès dû au seul hasard est d'autant plus grande que le nombre de points est faible, ce qui est notre cas.

Cependant, pour tout de même s'assurer du bien fondé de nos expériences, nous avons construit le diagramme de dispersion de la figure 16, qui montre de façon éloquente que l'utilisation de l'analyse de Fisher est justifiée.

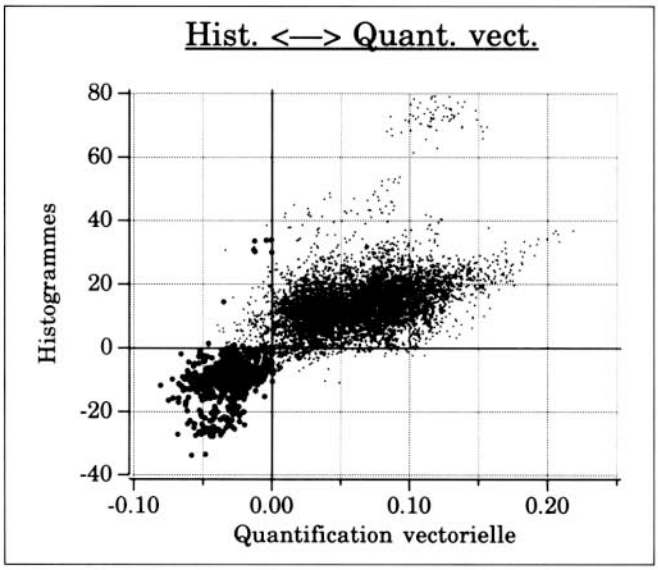


**Figure 16: Diagramme de dispersion des distances obtenues par quantification vectorielle et par histogrammes**

Nous voyons dans ce diagramme un nuage de points gras qui représente le domaine intra-locuteur (toutes références confondues); le domaine inter-locuteur est donné par les points maigres. Constatons que les distances obtenues par une méthode et par l'autre ne sont pas corrélées entre elles, puisque le lieu des points du diagramme n'est pas une droite. Constatons encore que chaque méthode est efficace par elle-même. Ces deux prémices nous permettent de conclure qu'une méthode est capable de corriger les éventuelles erreurs commises par l'autre, ce qui justifie l'utilisation de la méthode du discriminant linéaire de Fisher.

## 7. Conclusions

Nous avons décrit quatre méthodes différentes de reconnaissance de locuteurs indépendante du texte. Une de ces

méthodes, baptisée méthode de l'histogramme des vecteurs choisis par quantification vectorielle dans un vocabulaire universel, est inédite. Nous avons montré, par une série d'expériences en largeur de bande téléphonique, que trois des quatre méthodes sont efficaces par elles-mêmes. Nous avons conservé notre nouvelle méthode; la méthode que nous avons exclue est la technique d'accumulation d'erreur de quantification vectorielle de cepstres différentiels. Après l'avoir rejetée, nous avons montré comment combiner les autres résultats partiels pour améliorer le résultat global. Nous en obtenons un taux d'erreur équitable résiduel de 0,2%, ce qui nous paraît favorable dans un mode de reconnaissance indépendante du texte.

L'exclusivité de l'intérêt qu'accordent aux cepstres les méthodes proposées laisse poindre l'espoir que des techniques différentes puissent accéder à des caractéristiques complémentaires du signal de parole, susceptibles de contenir une partie de l'identité du locuteur. Par exemple, nous pourrions tenir compte de l'écoulement du temps, puisque l'ordre d'apparition des cepstres instantanés nous a été jusqu'à maintenant indifférent. Ou encore, sachant que les cepstres ont été calculés au moyen d'une analyse LPC délivrant trois séries de paramètres (énergie, résidu, coefficients du filtre inverse), nous pourrions en exploiter plus que la seule représentation cepstrale du filtre inverse.

Nous nous proposons, pour nos futures recherches en reconnaissance du locuteur, de porter l'accent sur le résidu du signal vocal (parfois appelé signal d'excitation), de sorte à considérer un aspect supplémentaire permettant de déterminer l'identité d'un locuteur, tout en restant indépendant du texte.


## Remerciements

Nous remercions la Fondation Hasler-Werke pour son soutien à ce projet, ainsi que tous les locuteurs ayant accepté de prêter leur voix.


## Bibliographie


[1] B.S. Atal, *«Automatic Recognition of Speakers from Their Voices»*, Proc. IEEE, Vol. 64, No. 4, Apr. 1976, pp. 460–475

[2] J.B. Attili, M. Savic, J.P. Campbell Jr., *«A TMS 3220-Based Real Time, Text-Independent, Automatic Speaker Verification System»*, ICASSP 88, New York City, pp. 599–602

[3] C. Bernasconi, *«Erhöhung der Entscheidungssicherheit textabhängiger Sprecherverifikationsverfahren»*, Mitteilungen AGEN, No. 49, Mai 1989, pp. 17–27

[4] G.R. Doddington, *«Speaker Recognition–Identifying People by their Voices»*, Proc. IEEE, Vol. 73, No. 11, Nov. 1985, pp. 1651–1664

[5] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley-interscience publication, 1973

[6] S. Furui, *«Cepstral Analysis Technique for Automatic Speaker Verification»*, IEEE Trans. ASSP, Vol. 29, No. 2, Apr. 1981, pp. 254–272

[7] S. Furui, *«Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features»*, IEEE Trans. ASSP, Vol. 29, No. 3, Jun. 1981, pp. 342–350

[8] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, J. Wolf, *«Investigation of Text-Independent Speaker Identification Over Telephone Channels»*, ICASSP 85, Tampa, pp. 379–382

[9] H. Gish, M. Krasner, W. Russel, J. Wolf, *«Methods and Experiments for Text-Independent Speaker Recognition Over Telephone Channels»*, ICASSP 86, Tokyo, pp. 865–868

[10] A.L. Higgins, R.E. Wohlford, *«A New Method of Text-Independent Speaker Recognition»*, ICASSP 86, Tokyo, pp. 869–872

[11] M.J. Hunt, *«Further Experiments in Text-Independent Speaker Recognition Over Communication Channels»*, ICASSP 83, Boston, pp. 563–566

[12] K.P. Li, J.E. Porter, *«Normalizations and Selection of Speech Segments for Speaker Recognition Scoring»*, ICASSP 88, New York City, pp. 595–598

[13] K.P. Li, E.H. Wrench Jr., *«An Approach to Text-Independent Speaker Recognition with Short Utterances»*, ICASSP 83, Boston, pp. 555–558

[14] J. Makhoul, S. Roucos, H. Gish, *«Vector Quantization in Speech Coding»*, Proc. IEEE, Vol. 73, No. 11, Nov. 1985, pp. 1551–1588

[15] N. Mohankrishnan, M. Shridhar, M. A. Sid-Ahmed, *«A Composite Scheme for Text-Independent Speaker Recognition»*, ICASSP 82, Paris, pp. 1653–1656

[16] J.M. Naik, G.R. Doddington, *«High Peformance Speaker Verification Using Principal Components»*, ICASSP 86, Tokyo, pp. 881–884

[17] H. Ney, R. Gierloff, *«Speaker Recognition Using a Feature Weighting Technique»*, ICASSP 82, Paris, pp. 1645–1648

[18] A.E. Rosenberg, F.K. Soong, *«Evaluation of a Vector Quantization Talker Recognition System in Text-Independent and Text-Dependent Modes»*, ICASSP 86, Tokyo, pp. 873–876

[19] M. Shridhar, N. Mohankrishnan, *«Text-Independent Speaker Recognition: A Review and Some New Results»*, Speech Comm., Vol. 1, No. 3–4, Dec. 1982, pp. 257–267

[20] M. Shridhar, N. Mohankrishnan, M. Baraniecki, *«Text-Independent Speaker Recognition Using Orthogonal Linear Prediction»*, ICASSP 81, Atlanta, pp. 197–200

[21] M. Shridhar, N. Mohankrishnan, M. A. Sid-Ahmed, *«A Comparison of Distance Measures for Text-Independent Speaker Identification»*, ICASSP 83, Boston, pp. 559–562

[22] R. Schwartz, S. Roucos, M. Berouti, *«The Application of Probability Density Estimation to Text-Independent Speaker Identification»*, ICASSP 82, Paris, pp. 1649–1652

[23] F.K. Soong, A.E. Rosenberg, *«On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition»*, ICASSP 86, Tokyo, pp. 877–880

[24] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, B.H. Juang, *«A Vector Quantization Approach to Speaker Recognition»*, ICASSP 85, Tampa, pp. 387–390

[25] P. Thévenaz, *«Combining Four Text Independent Speaker Recognition Methods»*, Proc. ESCA Research Workshop on Speaker Characterization in Speech Technology, Edinburgh, 1990, pp. 187–191.

[26] G. Velius, *«Variants of Cepstral Based Speaker Identity Verification»*, ICASSP 88, New York City, pp. 583–586

[27] J.J. Wolf, *«Efficient Acoustic Parameters for Speaker Recognition»*, The Journal of the Acoustical Society of America, Vol. 51, No. 6, Part 2, 1972, pp. 2044–2056

[28] L. Xu, J. Oglesby, J.S. Mason, *«The Optimization of Perceptually-Based Features for Speaker Recognition»*, ICASSP 89, Glasgow, pp. 520–523


## Auteur


Philippe Thévenaz
ing. dipl. EPFL

Institut de microtechnique
Université de Neuchâtel
A.-L. Breguet 2

CH–2000 Neuchâtel
Suisse

tél. 038/20 54 57
fax 038/25 42 76