RECOMMANDATION UIT-R BS.1657

Procédure de test des systèmes automatiques d'identification audio

(Question UIT-R 8/6)

(2003)

L'Assemblée des radiocommunications de l'UIT,

considérant

a) que, à terme, des métadonnées accompagneront la plupart des systèmes de diffusion audio;

b) que la production automatique de métadonnées sera nécessaire pour offrir dans l'avenir un service complet présentant un bon rapport coût-efficacité;

c) que l'identification automatique des séquences audio permet de suivre les programmes transmis;

d) que différents systèmes d'extraction de métadonnées sont mis au point actuellement;

e) que le GT de l'ISO/CEI JTC 1/SC 29 élabore actuellement, sous la forme définitive, des systèmes de codage de métadonnées pour les données multimédias;

f) que, jusqu'à présent, aucune procédure d'évaluation de la qualité des systèmes d'extraction de métadonnées audio n'a été normalisée,

recommande

1 d'utiliser la procédure décrite dans l'Annexe 1 pour évaluer la qualité de fonctionnement des systèmes automatiques d'identification audio.

Annexe 1

Procédure de test des systèmes automatiques d'identification audio

1 Introduction

A l'heure d'un accroissement toujours plus grand des bases de données à contenu musical, qu'elles contiennent de véritables données audio ou des métadonnées associées («données sur les données»), l'exigence d'outils permettant de conserver ces masses de données devient également chaque jour plus urgente. Ce souhait n'est pas seulement exprimé par des professionnels, mais également par le simple amateur de musique utilisateur de l'Internet qui navigue fréquemment sur la Toile à la recherche de son style musical préféré. Pour faciliter l'extraction des données souhaitées, on distingue ici deux niveaux d'abstraction:

– recherche de métadonnées plus ou moins susceptibles d'être extraites automatiquement du contenu audio (instruments, thème mélodique, rythme, etc.). Un système de requête par fredonnement ou de classification par genres, couramment utilisé par les moteurs de recherche, pourrait en constituer un exemple d'application;

– identification automatique des titres, lorsqu'on dispose seulement de métadonnées insuffisantes, non fiables ou lorsque aucune métadonnée n'est disponible. Une «esquisse» de données audio est générée puis comparée à une base de données connues, créant ainsi un lien vers des métadonnées pertinentes telles qu'un nom d'artiste, le titre d'une chanson, etc.

Si la première classe se rapporte essentiellement à l'interface avec l'homme, la seconde trouve également une application dans la protection des droits par la reconnaissance de programmes radiophoniques et de transactions Internet. Il est primordial dans ce dernier cas de souligner que les algorithmes appartenant à cette seconde classe sont désignés sous le terme de techniques de «prise d'empreintes audio».

2 Objet

Pour répondre aux exigences de l'industrie musicale, le taux de reconnaissance des techniques de reconnaissance d'empreintes audio appliquées doit être élevé et ne pas être dégradé par les altérations et modifications courantes subies par le contenu audio original. A cette fin, l'industrie musicale a reconnu la nécessité d'une garantie de qualité pour les systèmes d'identification audio en formulant récemment une demande d'informations sur les techniques de reconnaissance d'empreintes audio.

Le caractère crucial et urgent de ce problème est encore renforcé par le fait qu'un certain nombre de solutions différentes, souvent propriétaires, sont apparues récemment. Pour toutes ces méthodes cependant, les mêmes problèmes se posent quant à leur robustesse vis-à-vis de modifications ou de détériorations des données d'origine. Bien qu'il puisse avoir été modifié par un certain nombre d'étapes de traitement ou de dégradations, le contenu d'origine doit pouvoir être reconnu comme étant la propriété intellectuelle de l'artiste ou du compositeur. Il convient donc de proposer que l'identification automatique des données musicales soit idéalement aussi précise et robuste vis-à-vis de modifications apportées aux signaux que le sont la perception et la reconnaissance humaine. Au-delà de la robustesse vis-à-vis des altérations des signaux, un bon système de reconnaissance des empreintes audio devrait présenter une petite taille d'empreinte (compte tenu du fait que certaines applications pourraient nécessiter le stockage de millions d'empreintes), une extraction et une reconnaissance rapides des empreintes ainsi que d'autres propriétés souhaitables. Il convient de noter que la robustesse aux altérations des signaux et la compacité de la représentation des empreintes sont deux spécifications contradictoires, que ces systèmes doivent concilier.

En conséquence, pour évaluer la qualité d'un système automatique d'identification audio, il faut définir un environnement de test couvrant différents types de dégradation des signaux pour plusieurs degrés de gravité et décrivant la façon de déterminer d'autres paramètres essentiels du système. Une procédure de test unifiée est nécessaire pour parvenir à une évaluation objective des systèmes d'identification.

3 Paramètres de qualité

Il convient de considérer les paramètres de qualité ci-après pour les systèmes d'identification audio:

- Taille du segment de données audio à identifier:
  - quelle partie d'un enregistrement est nécessaire pour l'identification?
- Taille de l'empreinte audio:
  - combien de données (octets) par enregistrement doivent être stockées dans la base de données?
  - la taille de l'empreinte audio est-elle constante ou variable (par rapport à la durée de l'enregistrement)?
- Taille de la base de données:
  - combien d'enregistrements peuvent être traités simultanément par le système?

- Mode d'identification:
  - le système permet-il d'identifier des fragments de contenus audio choisis au hasard (prise d'empreintes audio continue) ou l'identification est-elle restreinte à des petits segments comportant des empreintes? Dans ce dernier cas, quelle est la taille de ce segment?
- Vitesse d'identification:
  - quel est le temps nécessaire pour identifier un enregistrement?
  - comment varie cette durée suivant le nombre d'enregistrements présents dans la base de données?
- Qualité de l'identification pour les données d'origine et les données altérées:
  - quel niveau de distorsion peut-il être introduit sans dégradation significative du taux de reconnaissance?
  - comment varie ce niveau de distorsion suivant le nombre d'enregistrements dans la base de données et le niveau réel de distorsion?
- Vitesse de génération des empreintes audio:
  - à quelle vitesse une empreinte audio peut-elle être générée sur une plate-forme donnée?
  - quelles sont les ressources nécessaires pour générer une empreinte audio (fréquence fonctionnement de l'unité centrale, quantité de RAM, unité de traitement à virgule flottante, par exemple)?
- Vitesse d'acquisition:
  - quel est le temps nécessaire pour ajouter des enregistrements dans la base de données? Comment varie cette durée suivant le nombre d'enregistrements déjà présents dans la base de données?

Pour évaluer ces propriétés d'une manière réaliste et donc pour déterminer si un système est adapté à des applications réelles, un environnement de test doit présenter des conditions aux limites constantes en ce qui concerne les caractéristiques testées.

Les conditions de test doivent porter sur la taille et le contenu de la base de données de référence, sur la taille (en termes de durée d'enregistrement) et le nombre d'enregistrements de test, sur les règles exactes de modification des enregistrements de test et sur la plate-forme de calcul (spécification de l'unité centrale, de la mémoire et du système d'exploitation). Il convient également d'ajouter à l'ensemble des enregistrements de test un certain nombre de titres ne figurant pas dans la base de données de référence, afin d'évaluer de manière appropriée les caractéristiques de rejet du système testé.

4 Sélection des données de test et taille de la base de données

L'ensemble des différents styles et genres musicaux devrait être présent dans la base de données de référence, en accordant une place privilégiée aux genres les plus entendus. Une base de données de 10 000 à 100 000 titres est suggérée pour une estimation réaliste.

Définition des termes:

- On parle d'enregistrement dupliqué par rapport à un autre enregistrement audio s'il s'agit d'un enregistrement identique à l'original à l'exception éventuelle d'un certain nombre de zéros ajoutés en début ou en fin d'enregistrement. On peut parfois constater ce cas lorsque la «même» chanson figure sur des compilations ou albums différents.
- Un enregistrement similaire est un (re)mixage, une reprise ou un enregistrement (en public) d'un autre élément de la base de données.

Spécifications relatives à la sélection des données de test:

- Il faudrait tout particulièrement éviter la présence d'enregistrements dupliqués dans la base de données.
- La base de données devra contenir un certain nombre d'enregistrements similaires (20 paires au minimum). Par exemple:
  - 10 enregistrements en public par un artiste de la même chanson lors de différents concerts;
  - 10 paires enregistrement original/enregistrement remixé d'une même chanson par des artistes différents;
  - 10 paires enregistrement original/reprise d'une même chanson par des artistes différents.
- La base de données devra être définie avant de procéder au premier test. Il n'est pas permis de la modifier en fonction des résultats des tests.

5 Méthode de test

La rapidité des calculs risquant de dépendre du niveau de distorsion de l'enregistrement de test, il est nécessaire de mesurer séparément pour chaque expérience (Tests 1, 2, 3a) à 3i)) la vitesse du processus d'extraction et de recherche (classification).

5.1 Test 1

Lors du premier test, aucun titre de la base de données de référence ne doit être modifié et tous les titres doivent être identifiés. Le système testé devrait donc afficher un taux d'identification correct des enregistrements égal à 100%.

La taille moyenne d'une empreinte audio est calculée sur la base de la totalité des enregistrements de référence, d'où une taille moyenne par enregistrement ou une taille par durée de l'enregistrement dépendant du type d'empreinte audio du système testé. Il faudra considérer séparément les données issues de systèmes ne permettant pas la prise d'empreintes audio continue et les données issues de systèmes permettant ce type de prise d'empreintes.

5.2 Test 2

Des extraits de 1 000 enregistrements de 5 s et de 1 000 enregistrements de 30 s ne figurant pas dans la base de données de référence, et donc inconnus du système, devront ensuite être ajoutés à l'ensemble des données de test. Ces 2 000 extraits sont présentés au système afin d'évaluer ses caractéristiques de rejet et de tester le risque de reconnaissances signalées à tort. Au moins 10 de ces 2 000 enregistrements devraient être du type «enregistrement similaire» (à un enregistrement correspondant dans la base de données de référence).

5.3 Test 3

Pour tester la robustesse vis-à-vis de la modification de titres musicaux, on choisit un ensemble de 1 000 enregistrements dans la base de référence. Le premier test doit être effectué conformément aux descriptions du point 3a). Tous les autres tests (3b) à 3i)) sont ensuite basés sur les extraits créés au 3a), ce qui signifie qu'ils associent à l'effet de «recadrage» décrit à ce point une distorsion particulière. Le fait d'associer au recadrage toutes les autres distorsions paraît raisonnable pour ne pas supposer implicitement que les empreintes audio sont parfaitement homogènes, ce qui serait irréaliste.

Il est recommandé d'utiliser les procédures de modifications suivantes:

3a) Recadrage/décalage

On ne prend que des sous-segments de petite taille de l'enregistrement de test. Le choix du premier échantillon de l'extrait est indifférent (échantillon choisi aléatoirement mais le même pour tous les systèmes de test). La durée de l'extrait devrait être de 5, 10 ou 20 s.

3b) Compression et extension dynamiques

Les paramètres doivent être choisis en fonction des paramètres habituels utilisés pour la radiodiffusion.

3c) Réglage du niveau

Appliquer au signal d'entrée un certain facteur d'échelle (–6 dB et 10 dB, par exemple). L'écrêtage devra être évité.

3d) Egalisation

Utilisation d'une égalisation par octave avec des affaiblissements dans les bandes adjacentes fixés à –6 dB et +6 dB.

3e) Addition de bruit

Addition de bruit blanc ou rose avec une valeur globale de *S*/*N* égale respectivement à 10 ou 20 dB.

3f) Conversion du taux d'échantillonnage et changement de hauteur.

Des déviations de +5% et –5% du taux d'échantillonnage doivent être utilisées.

3g) Codage audio et «tatouage numérique»

L'incidence d'un codage audio doit être évaluée en utilisant un signal codé MPEG-1/2 de couche 3 présentant les associations débit binaire/canal suivantes: 24 kbit/s (mono), 64 kbit/s (stéréo), 96 kbit/s (stéréo) et 128 kbit/s (stéréo).

3h) Limitation de la bande

La borne supérieure de la bande passante du signal d'entrée doit être limitée à 4 kHz.

3i) Transmission acoustique

Les imperfections causées par un retour sonore en conditions acoustiques moyennes doivent être testées: le signal est émis par un haut-parleur puis réenregistré par un microphone. La distance recommandée entre ces deux dispositifs est d'environ 50 cm. Il est inutile de choisir un haut-parleur et/ou un microphone de qualité supérieure. Ce test devrait être effectué dans une pièce ordinaire (sans traitement acoustique ni isolation).

Les paramètres des différents tests de modification ont été réglés de telle manière qu'une perception d'écoute humaine équivalente qualifierait d'«altération légère» à «altération forte» les modifications apportées aux données d'origine. Pour le codage audio, l'altération légère correspondrait à un codage MP3 à 128 kbit/s (stéréo) et l'altération forte à un codage MP3 à 24 kbit/s. Des codages intermédiaires à 96 kbit/s (stéréo) et 64 kbit/s (stéréo) sont recommandés, étant donné que ces débits sont le plus couramment utilisés pour les transactions Internet. Il convient donc de choisir un maximum de 5 niveaux de dégradation[1].

---

[1] On considère que l'inclusion des codes MPEG-1/2 de couche 2, MPEG-2/4 AAC, Dolby-E ou autres, fréquemment utilisés dans les environnements de radiodiffusion, n'est pas nécessaire parce que ces algorithmes ne sont généralement pas «mal utilisés» dans un environnement d'étude, contrairement à ce qui se produit fréquemment pour le codage MPEG-1/2 de couche 3 (MP3).

6 Plate-forme de test

Il convient d'utiliser comme plate-forme de calcul et système d'exploitation des équipements adaptés à l'état d'avancement des techniques offertes à l'utilisateur courant. En 2002, on peut citer comme plate-forme appropriée un ordinateur Pentium fonctionnant à 1 GHz avec 512 Mégaoctets de RAM et utilisant Windows 2000™ ou Linux.

7 Modification des paramètres du système

Au cours des différents tests, le réglage des systèmes de prise d'empreintes audio, qui permettent de parvenir à différents degrés de robustesse et de compacité de prises d'empreintes en fonction du réglage des paramètres d'extraction, peut être modifié en vue d'obtenir une qualité de fonctionnement optimale pour chaque fonction ou chaque test. Chaque association système/réglage de paramètres doit alors cependant être considérée comme un système indépendant ayant un champ d'application limité, son propre format de prises d'empreintes audio et son propre processus d'extraction. Ces considérations ne s'appliquent pas aux systèmes pour lesquels une base de données d'empreintes audio plus compacte ou moins robuste peut être obtenue à partir d'une représentation moins compacte ou plus robuste au moyen d'un processus de transcodage autonome, c'est-à-dire lorsqu'un seul processus d'extraction d'empreintes audio appliqué aux données audio de référence suffit pour effectuer toutes les fonctions intervenant dans les tests.

8 Rapport de test

Les rapports de test devraient indiquer, aussi clairement que possible, la logique de l'étude, les méthodes utilisées et les conclusions auxquelles on a abouti. Ils devraient être assez détaillés pour qu'une personne raisonnablement compétente puisse, en principe, reproduire l'étude afin d'en vérifier empiriquement les résultats. Un lecteur informé devrait être capable de comprendre les principaux détails du test et d'en développer un point de vue critique, concernant par exemple les raisons sous-jacentes ayant motivé l'étude, les méthodes de conception expérimentales et leur mise en oeuvre, ainsi que les analyses et les conclusions.

Une attention particulière devrait être portée aux points suivants:

- la spécification et la sélection des enregistrements de référence et des enregistrements de test;
- la sélection des enregistrements similaires et les résultats de test associés à ces enregistrements particuliers;
- la description détaillée des paramètres des différentes distorsions;
- la description détaillée des paramètres de réglage utilisés pour les systèmes testés;
- le détail des éléments de base ayant sous-tendu l'ensemble des conclusions auxquelles on a abouti.