SEGMENTACIÓ DE LOCUTOR PER A L'INDEXACIÓ AUTOMÀTICA DE BASES DE DADES MULTIMÈDIA EN CATALÀ

Xavier Anguera, Mireia Farrús, Javier Hernando
Departament de Teoria del Senyal i Comunicacions (TSC), Centre de Recerca TALP
Universitat Politècnica de Catalunya (UPC), Barcelona.
{xanguera, mfarrus, javier} @gps.tsc.upc.es

1. RESUM

L'evolució de la societat de la informació ha esdevingut un incessant increment de continguts audiovisuals que s'emeten constantment en cadenes de televisió i emissores de radio locals i nacionals en llengua catalana. Aquestes emissions normalment s'arxiven en bases de dades multimèdia per tal de poder ser consultades posteriorment, però degut a la gran quantitat de dades emmagatzemades resulta difícil, si no impossible, i molt costós poder accedir a aquesta informació.

Amb aquesta comunicació pretenem donar a conèixer les tècniques existents actualment d'indexació automàtica de material sonor en les quals estem treballant en el departament de Teoria del Senyal i Comunicacions de la UPC. Mitjançant una indexació automàtica de les bases de dades és possible realitzar cerques concretes i recuperar documents molt més ràpidament.

Mostrem especial èmfasi en el cas de la indexació de la identitat de les persones que apareixen a la base de dades, i en quin interval de temps parlen. Presentem una mesura anomenada XBIC per detectar els canvis de locutor dins d'un senyal de veu, creada dins del nostre grup. Es mostren resultats d'aquesta nova tècnica sobre una base de dades recollida en llengua catalana.

2. INTRODUCCIÓ

En l'actual societat de la informació en què ens trobem immersos són moltes les emissores de radio i de televisió que emeten els seus continguts 24 hores al dia, tots els dies de l'any. Aquests continguts s'acostumen a arxivar per poder ser utilitzats amb posterioritat en futurs programes, entre d'altres. Fins no fa gaire, l'única manera d'arxivar aquesta informació era mitjançant suport analògic com ara cintes d'àudio o vídeo, les quals omplien grans sales d'arxius que necessitaven condicions ambientals òptimes perquè els materials no es deterioressin.

Amb l'arribada dels suports digitals, tota aquesta programació es pot emmagatzemar ja en format digital, amb tres avantatges principals. El primer és la protecció de les dades, les quals no sofreixen cap deteriorament degut al material que les conté; el segon, el tractament directe de les dades amb equips d'edició digitals, totalment estesos avui en dia; i finalment, la possibilitat d'accés ràpid i còmode a les dades per més d'una persona a la vegada, fins i tot quan aquestes es troben en un punt físic diferent del de l'arxiu digital.

Per tal de poder accedir de manera eficient a la informació continguda en aquests arxius (tant en format analògic com digital) s'han desenvolupat sistemes d'indexació que permeten posteriorment una cerca ràpida dels fragments desitjats. El problema principal d'aquests mètodes d'indexació és que moltes vegades s'han de dur a terme de manera manual (etiquetes a les cintes, entrades a mà a una base de dades...) de manera que resulta molt costós mantenir un registre exhaustiu dels continguts de l'arxiu.

En aquesta publicació volem posar de manifest l'existència de tecnologies que permetrien la indexació automàtica d'aquests arxius per poder obtenir molta més informació dels continguts emmagatzemats. En ser automàtica, la indexació implica un estalvi en hores de feina manual i manté un criteri constant durant tot el procés. De la mateixa manera, es facilita la recuperació d'aquests continguts de manera molt més efectiva.

Els mètodes d'indexació automàtica comprenen tècniques de tractament d'àudio i vídeo. En aquesta publicació ens centrem en la part d'àudio, comuna en els continguts d'emissores de radio i de televisió.

3. INDEXACIÓ DEL SENYAL D'ÀUDIO

A l'hora d'emmagatzemar i indexar el senyal d'àudio de les emissions tant de radio com de televisió, cal tenir en compte que aquest està compost per molts tipus de continguts: veu, publicitat, silencis o soroll, entre d'altres. En els intervals on hi ha veu podem distingir entre una o més persones parlant, on es troba el locutor (estudi, carrer...) i la presència de possibles sons externs afegits, com ara música o altres veus de fons.

A diferència de la indexació manual, en indexar l'àudio automàticament podem incloure molta més informació sense un cost afegit important. Això s'assoleix concatenant diferents sistemes de tractament del senyal de veu especialitzats en l'extracció de característiques particulars. Normalment s'indexen de manera exhaustiva aquells fragments que són de veu, deixant només amb marques d'inici i final els altres tipus. Algunes de les informacions indexables són la identitat i el lloc on apareixen cadascuna de les persones que parlen, el que diuen els locutors i en quin ambient de soroll i en quines condicions (lectura, parla espontània) ho diuen.

De manera general podem veure en la figura 1 els blocs principals que un sistema d'indexació de veu hauria de tenir. Moltes de les tècniques utilitzades poden treballar a temps real, tot i que no és estrictament necessari per a la indexació, ja que els documents es tracten posteriorment a la seva emissió i la rapidesa dels algorismes no és crucial.

4. SEGMENTACIÓ DE LOCUTOR

Entenem per segmentació de locutor del senyal de veu el conjunt d'algorismes que permeten, donat un segment d'àudio, partir-lo en els diferents locutors que hi apareixen, definint els punts en què comença i acaba cadascun d'ells. Es poden definir tres nivells de profunditat en la segmentació: en un primer nivell (anomenat segmentació de locutor) el resultat del sistema és un conjunt de marques temporals que defineixen els diferents locutors, sense donar cap informació sobre la seva identitat o si apareixen un o més cops en el total de la gravació. En un segon nivell, que s'anomena agrupament de locutors (*speaker clustering*) es comparen els diferents segments entre si i s'assignen al locutor al qual pertanyen. Aquesta tasca es pot dur a terme tant dins d'una mateixa gravació (aglomerat intra-sessió) o entre diferents gravacions,

pertanyents per exemple a programes de diferents dies (aglomerat inter-sessió) [1]. Un tercer nivell de segmentació (anomenat identificació de locutor) consisteix a associar cada aglomerat de segments amb un locutor concret, conegut a priori.

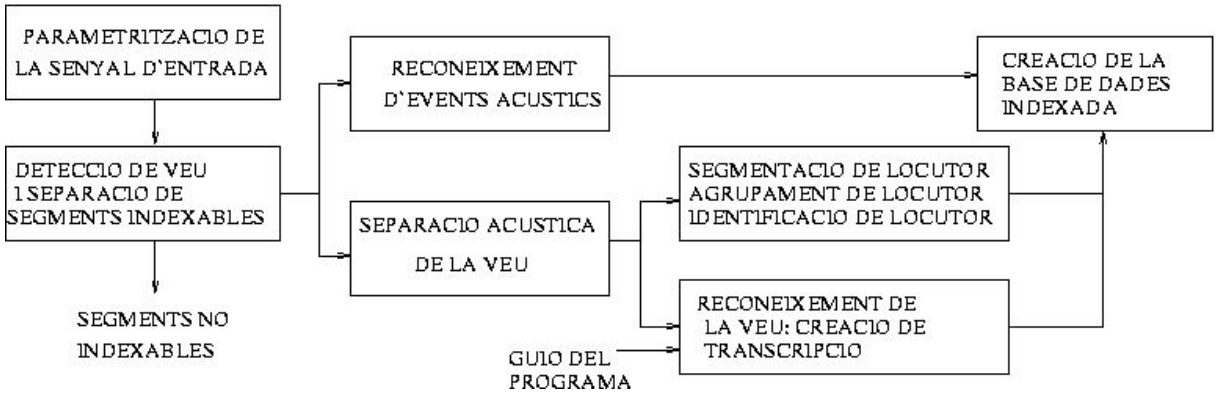

Figura 1: Blocs bàsics d'un sistema d'indexació de veu.

5. ALGORISME XBIC DE SEGMENTACIÓ DE LOCUTOR

Per a la segmentació d'un fragment de veu en els diferents locutors que hi intervenen presentem l'algorisme XBIC de mesura de distàncies entre dos segments de senyal. Donats dos segments de veu, l'algorisme aporta una mesura de similitud o distància entre aquests dos, de manera que podem considerar que els dos segments pertanyen a dos locutors diferents quan aquesta distància supera un llindar determinat. La mesura XBIC dóna un mínim en el punt on hi ha un canvi de locutor.

Per aplicar aquesta mesura s'utilitzen dues finestres lliscants per sobre del senyal, de longitud fixa i igual a totes dues, connectades per un punt de mesura que anomenarem x(i). El procés es fa en dues passades, una de ràpida i una altra de més refinada. Podem veure el funcionament de l'algorisme a la figura 2.

Els dos segments de longitud T es troben en el punt x(i), que és on es comprova si hi ha un canvi de locutor. En un primer pas, es calcula la mesura XBIC i es desplaça tot el conjunt cap endavant en intervals de T2 segons. Quan el valor XBIC compleix la condició de canvi de locutor es fa un segon pas dins d'un interval al voltant del punt detectat i amb un pas de desplaçament molt menor que abans, per trobar el punt exacte de canvi. Això es repeteix fins al final del senyal, cercant tots els canvis de locutor existents. Aquest procediment deixa sempre un espai de longitud T al principi i al final del segment de senyal avaluat on no es reconeix cap canvi de locutor.

El valor de T ha de ser prou gran com per a que la mesura XBIC es pugui dur a terme, però també el mínim possible perquè l'espai no avaluat al principi i final no suposi cap problema. En el nostre cas s'ha utilitzat un valor de 2 segons.

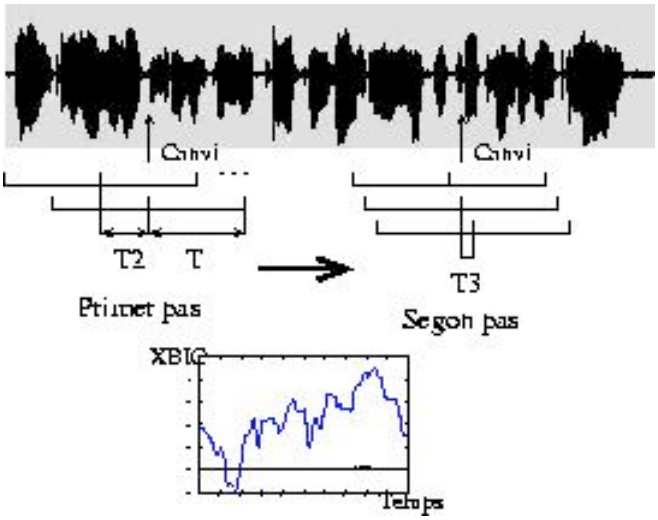

Figura 2: Segmentació de locutor en dues passades utilitzant XBIC

La mesura XBIC té certes similituds amb [2] i [3] i s'obté de la manera següent:

$$XBIC = P(A \mid \lambda_B) + P(B \mid \lambda_A)$$

on A i B son dos segments d'entrada de mateixa durada. Amb aquests segments s'entrenen els models de Markov [4] (anomenats λ_A i λ_B) i s'avalua amb el segment oposat respectiu.

6. PROVES DE SEGMENTACIÓ DE LOCUTORS AMB DADES EN CATALÀ

Per a l'avaluació del l'algorisme de segmentació XBIC s'han recollit i segmentat a mà 2,5 hores de senyal d'àudio procedents de diferents transmissions del Telenotícies de les cadenes de TV3 (3 telenotícies migdia) i 3/24 (1 Telenotícies matí). En tots els casos s'han eliminat els anuncis emesos enmig d'aquestes emissions.

Els senyals s'han desat en format PCM, amb 16 bits/mostra i a una freqüència de mostreig de 16 KHz. Per ser tractats pel sistema de segmentació, s'han parametritzat utilitzant paràmetres MFCC de 16 mostres + 16 derivades de primer ordre.

L'algorisme de segmentació s'ha aplicat en tot el senyal però per a l'avaluació només s'ha tingut en compte la segmentació resultant en les àrees on apareixen locutors. Els models de Markov utilitzats es formen amb 1 estat compost per 1 gaussiana de matriu de covariances plena (tots els valors són diferents de 0). Per a cada iteració els models s'entrenen mitjançant entrenament EM (Expectation Maximization). La mètrica és la utilitzada en aquest tipus de segmentació, tenint en compte dos tipus d'error:

- Mesurem l'error degut al fet de trobar més canvis del compte amb la mesura PRC
  PRC = Nombre de canvis trobats correctament / nombre total de canvis trobats
- Mesurem l'error degut al fet de no trobar els canvis amb la mesura RCL
  RCL = Nombre de canvis trobats correctament / nombre total de canvis existents.

En el cas de funcionament òptim les dues mesures arribaran al 100%, i en el pitjor cas baixaran fins al 0%. A la taula següent podem veure els resultats:

Telenotícies	*Mesura RCL*	*Mesura PRC*
TV3 migdia 6-7-2004	50.00%	67.00%
TV3 migdia 7-7-2004	43.01%	63.40%
TV3 migdia 8-7-2004	53.04%	64.89%
3/24 TN matí 5-7-2004	46.47%	41.27%
Mitjanes:	48.13%	59.14%

L'anàlisi dels resultats permet veure que hi ha tres tipus principals de problemes que causen que el sistema de segmentació produeixi errors:

- durant els períodes de transmissió on hi ha locutors parlant amb música de fons (com en la lectura dels titulars) el sistema tendeix a confondre els locutors en no distingir entre la música i la veu.
- quan un locutor està parlant es poden produir canvis en la transmissió de vídeo que provoquen canvis en el soroll ambiental, produint possibles insercions errònies de canvis, o bé quan un locutor canvia de parla llegida a espontània.
- quan hi ha vàries persones parlant alhora (com en el cas de les traduccions) el sistema tendeix a confondre's entre les dues.

7. CONCLUSIONS I TREBALL FUTUR

En aquesta publicació hem presentat una manera d'abordar el problema de l'indexació dels documents que constantment es generen per part de cadenes de radio i televisió a Catalunya. Actualment hi ha grans equips de persones realitzant una tasca limitada d'indexació per a un possible ús posterior. Mitjançant sistemes d'indexació automàtica d'àudio i vídeo es poden aconseguir bons resultats que permetrien ajudar a fer aquesta tasca molt més fàcil i amb una indexació molt més rica dels continguts.

Dins del grup de veu i imatge del departament de Teoria del Senyal i Comunicacions de la Universitat Politècnica de Catalunya ja es treballa en molts dels sistemes que permetrien dur a terme aquesta tasca. Tot i això, encara hi ha feina a fer per posar en funcionament aquests sistemes conjuntament i dur a terme una indexació complerta. El sistema de segmentació que presentem aquí té encara un percentatge d'error molt millorable però pretén donar a conèixer la recerca que es porta a terme en aquest àmbit i el fet que un sistema d'aquestes característiques pot ser factible i molt útil quan s'aplica sobre continguts en llengua catalana.

8. BIBLIOGRAFIA

[1] S.E. Tranter, D.A. Reynolds, "Speaker Diarization for Broadcast News", Comunicacions de Odyssey-2004, pp 337-344, Toledo, Juny 2004.

[2] J. Ajmera, I. McCowan, H, Bourlard, "Robust Speaker Change Detection", IDIAP Research Report 02-39, Setembre 2003.

[3] B.H. Juang, L.R.Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models", AT&T Technical Journal, vol 64, No 2, Febrer 1985.

[4] L. Rabiner, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, vol 77, No 2, pp 257-286, Febrer 1989.