LA PROSÒDIA COM A IDENTIFICADOR BIOMÈTRIC

Mireia Farrús i Cabeceran
Departament de Tecnologia Educativa
Universitat Oberta de Catalunya

1. QUÈ ÉS LA BIOMETRIA?

La biometria —o reconeixement biomètric— és la utilització de característiques distintives per reconèixer la identitat dels éssers humans (Maltoni et al. 2003, Bolle et al. 2004); una idea que, per primera vegada ara fa més d'un segle, Alphonse Bertillon va concebre i posar en pràctica per tal de resoldre crims (Jain et al. 1994).

Aquestes característiques biomètriques poden ser o bé fisiològiques, és a dir, característiques físiques com les empremtes dactilars, la cara, la geometria de la mà, la retina, l'iris, etc. que es poden mesurar en un determinat moment del temps, o bé de conducta, com la signatura, la veu o la manera de caminar, que consisteixen en com una acció es desenvolupa en un espai de temps. A diferència de les característiques fisiològiques, les característiques de conducta s'aprenen en el temps i es poden modificar fàcilment de manera deliberada (Bolle et al. 2004).

No obstant això, tot i que la classificació de característiques biomètriques entre identificadors fisiològics i de conducta sembla, a priori, molt clara, la majoria d'identificadors són la combinació de totes dues característiques. Els identificadors de conducta estan fortament relacionats amb el moviment, però depenen, en un grau molt alt, de l'estructura fisiològica de l'individu. La veu i el pas (la manera de caminar), per exemple, depenen de l'estructura anatòmica del mecanisme vocal i de les cames, respectivament (Brand et al. 2001).

Per saber quines característiques humanes són aptes per utilitzar-les com a identificadors biomètrics, Maltoni et al. (2003) proposen que aquestes característiques compleixin els requisits següents:

a. Universalitat: totes i cadascuna de les persones haurien de tenir l'identificador en qüestió.

b. Unicitat: l'identificador de qualsevol parell d'individus hauria de permetre distingir-los.

c. Permanència: l'identificador hauria de ser prou invariant al llarg del temps.

d. Facilitat d'ús: la mostra biomètrica hauria de ser fàcil d'adquirir.

No obstant això, a la pràctica hi ha altres factors que caldria tenir en compte en el funcionament dels sistemes biomètrics, com ara la *precisió*, la *intrusivitat* (contacte no desitjat amb l'individu a l'hora d'adquirir les dades biomètriques) o *l'acceptació social* (fins a quin punt els usuaris acceptarien un determinat identificador biomètric en el seu dia a dia).

L'objectiu d'aquest article és, en primer lloc, donar una visió general del paper de la prosòdia dins del reconeixement del locutor (seccions 2 i 3), i en segon lloc, demostrar en una part més experimental (secció 4) que la prosòdia té un paper important en el reconeixement automàtic del locutor.

2. La veu en el reconeixement biomètric de persones

La veu és un dels identificadors biomètrics més habituals per la seva poca intrusivitat i gran acceptació per part dels usuaris, tot i que, segons la percepció de Maltoni et al. (2003), també es caracteritza per una universalitat i facilitat d'ús mitjanes, i per una unicitat i permanència baixes. La qualitat del senyal de veu es pot degradar a través del micròfon o el canal de transmissió; la mateixa veu es pot veure afectada per diversos factors com la salut, l'estrès o les emocions i, a més a més, s'ha pogut demostrar que algunes persones tenen un talent extraordinari a l'hora d'imitar les veus dels altres (Maltoni et al. 2003, Bolle et al. 2004). Per aquest motiu, la majoria de sistemes biomètrics basats únicament en la utilització de la veu —d'ara en endavant, sistemes de *reconeixement automàtic del locutor* (RAL)— encara requereixen millores pel que fa a la fiabilitat.

Una de les qüestions centrals en l'àmbit de la recerca en el RAL és intentar esbrinar què és el que fa que la veu proporcioni signes d'identitat en un individu. Nosaltres, els humans, ens fixem en diferents nivells d'informació continguts en el senyal de veu a l'hora d'identificar les persones (Schmidt-Nielsen i Crystal 2000). Aquests nivells es poden relacionar amb diferents aspectes de la veu com el que els parlants anomenen popularment timbre, una riallada característica o la repetició de paraules específiques.

Tradicionalment, aquests nivells d'informació s'han classificat d'una manera jeràrquica, i s'acostuma a parlar d'informació de baix nivell, associada fonamentalment a aquelles característiques relacionades amb els trets físics de l'aparell vocal, i d'informació d'alt nivell, associada majoritàriament a aquelles característiques que depenen dels hàbits apresos i de l'estil de la parla —com la prosòdia o l'ús d'un vocabulari particular.

Fins fa pocs anys, la gran majoria de sistemes de RAL s'ha basat fonamentalment en característiques de baix nivell relacionades amb l'espectre de la veu. No obstant això, hi ha altres nivells d'informació que tenen un paper importantíssim en el procés de reconeixement del locutor per part dels humans, la qual cosa ens fa pensar que també contenen informació útil per al RAL. Els estudis que demostren que aquests nivells d'informació poden afegir informació complementària i millorar la precisió dels sistemes (Doddington 2001, Andrews et al. 2002, Bartkova et al. 2002, Weber et al. 2002, Peskin et al. 2003, Reynolds et al. 2003, Adami 2007, Cicres 2007) són cada vegada més nombrosos.

3. La quantificació dels elements prosòdics

La prosòdia es transmet a partir de tres elements diferents: l'entonació, la durada i la intensitat. Els éssers humans utilitzen diversos mètodes per produir aquests fenòmens lingüístics que determinen la prosòdia. Els canvis en el sistema respiratori i en els músculs de la laringe, per exemple, tenen un paper important a l'hora de controlar la freqüència fonamental (Atkinson 1978).

D'una banda, la freqüència fonamental ve determinada fisiològicament pel nombre de cicles que les cordes vocals produeixen en un segon, i per la seva tensió, i són el resultat natural de la longitud d'aquestes estructures. D' altra banda, la intensitat vocal es relaciona directament amb la pressió subglòtica de la columna d'aire. Aquesta pressió subglòtica, alhora, depèn de diversos factors com l'amplitud de la vibració i la tensió de les cordes vocals.

Però tots dos elements prosòdics es

relacionen mútuament; les variacions de la intensitat poden estar relacionades, en un grau més o menys elevat, amb la freqüència, ja que l'increment del to de la laringe genera una pressió glòtica més elevada i, en conseqüència, una intensitat més elevada. Per aquest motiu, les veus més agudes tendeixen a anar associades a una intensitat més elevada.

La prosòdia té un paper molt important en els actes de parla i la comunicació en el dia a dia. A la vegada, la parla s'ajusta a les particularitats del parlant, de manera que cada persona pot utilitzar diferents variacions de to, intensitat i ritme. Aquests canvis, que es manifesten en la prosòdia, són essencials per proporcionar una determinada melodia a la parla (Wertzner et al. 2005).

En un intent de quantificar els elements prosòdics de la parla per tal d'extreure'ls del senyal de veu i utilitzar-los en aplicacions biomètriques, es va dur a terme, en el *2002 Summer Workshop on Language Engineering* a la Johns Hopkins University, el projecte *SuperSID*, amb l'objectiu de millorar la fiabilitat de la tecnologia del RAL, explorant noves fonts d'informació en la parla conversacional (Peskin et al. 2003, Reynolds et al. 2003). En aquest projecte es va explorar i desenvolupar l'extracció i el modelatge dels patrons específics dels locutors en senyals acústics, la prosòdia de la parla, les pronúncies de les paraules i fonemes, l'ús idiomàtic de la llengua i les interaccions amb altres locutors.

En el marc del *SuperSID*, Peskin et al. (2003) es van centrar en investigacions sobre diverses recopilacions de característiques prosòdiques a partir del corpus de parla conversacional *Switchboard-I*. Com a banc de proves de referència es va utilitzar el protocol establert de l'*Extended Data Task* de la NIST (National Institute of Standards and Technology) del 2001, [1] basat en el corpus *Switchboard-I* per l'àmplia varietat de recursos disponibles per a aquest corpus (Reynolds et al. 2003).

Dels treballs de Shriberg et al. (2000) i Peskin et al. (2003) se n'extreu que algunes de les característiques prosòdiques que poden resultar útils per a la identificació biomètrica són les següents:

a. Logaritme del nombre de trames per paraula
b. Fracció de trames sonores en cada paraula
c. Logaritme de la F0 mitjana
d. Logaritme de la F0 màxima
e. Logaritme de la F0 mínima
f. Logaritme del rang de la F0 (F0màx.– F0mín.)
g. Pendent de F0 (últimaF0-primeraF0) / trames)
h. Pendent del contorn estilitzat de la F0
i. Freqüència relativa de les pauses curtes (7-15 trames)
j. Freqüència relativa de les pauses mitjanes (16-99 trames)
k. Freqüència relativa de les pauses llargues (més de 100 trames)
l. Logaritme de la longitud de les pauses curtes
m. Logaritme de les pauses mitjanes
n. Logaritme de les pauses llargues

Les característiques finals s'obtenen mitjançant el càlcul de la mitjana de cadascuna d'aquestes característiques per a totes les paraules o pauses de la conversa. En les característiques *i*, *j* i *k*, relacionades amb la freqüència relativa de les pauses, les trames (en aquest experiment) són segments de 10 milisegons de longitud.

[1] http://www.itl.nist.gov/iad/mig/tests/spk/2001/extended_data.html

4. La prosòdia en el reconeixement automàtic del locutor

Aquesta secció vol mostrar els avantatges d'utilitzar la prosòdia en la tasca del RAL; és a dir, com la combinació de la informació prosòdica i espectral permet millorar el funcionament global d'un sistema de reconeixement. Per a tal fi, ens remetrem a un dels experiments desenvolupats a Farrús (2008).

4.1. Conceptes previs

Tot i que no entrarem en gaire detall a l'hora de descriure els experiments desenvolupats, creiem necessari introduir breument els conceptes següents per poder entendre millor la metodologia utilitzada. No obstant això, per als lectors amb més formació filològica poc familiaritzats amb els conceptes tècnics que s'hi exposen, es recomana la lectura d'Hernando (2001).

4.1.1. Identificació i verificació

En funció del tipus d'aplicació, els sistemes de reconeixement biomètric poden funcionar en dos modes diferents: *identificació* o *verificació*. L'objectiu de la identificació és determinar quin parlant, d'entre un conjunt de parlants els models dels quals estan emmagatzemats en una base de dades, coincideix amb un usuari desconegut[2].

En la verificació, en canvi, es tracta de determinar si un usuari resulta ser qui diu que és. Aquest tipus de reconeixement és propi de les aplicacions relacionades amb restriccions d'accés en àrees de seguretat. La base de dades corresponent ha d'incloure un model d'impostor i un model corresponent a la identitat que l'usuari reclama (el client). Les característiques biomètriques d'aquest usuari es comparen amb els models del client i de l'impostor, i s'associen al model més similar.

La fiabilitat d'aquests sistemes es mesura en termes d'*Equal Error Rate* (EER), que és el valor en què la *taxa de fals positiu* (percentatge de correspondències errònies) i la *taxa de fals negatiu* (percentatge de correspondències vàlides rebutjades) són equivalents. Com més baix és el valor de l'EER, més fiabilitat té el sistema.

4.1.2. Normalització i fusió

Quan s'utilitza un identificador en la tasca de reconeixement biomètric, s'obtenen una sèrie de valors que indiquen la similitud d'aquest identificador amb un patró de referència al qual s'està fent la comparació. Com que cada identificador utilitza la seva pròpia escala, abans de combinar aquests valors cal normalitzar-los. Una de les tècniques més habituals i més senzilles és la *min-max*, que consisteix a escalar tots els valors en un rang que va del 0 a l'1, de manera que després es poden combinar —fusionar— fàcilment.

D'entre les tècniques de fusió, una de les més utilitzades és el *matcher weighting*, en què les mesures obtingudes per cada identificador biomètric es ponderen per un factor proporcional a la seva taxa de reconeixement, de manera que els pesos dels identificadors més fiables són més alts que els dels menys fiables (Indovina et al. 2003).

4.2. Base de dades

Per als experiments de reconeixement del locutor —concretament de verificació—que s'han fet servir en aquesta secció, s'ha utilitzat la base de dades *Switchboard-I* (Godfrey et al. 1990), una col·lecció de 2.430 converses telefòniques enregistrades per parelles i amb un total de 543 parlants diferents de totes les àrees dels Estats Units.

[2] Per identificar un parlant (locutor), primer cal entrenar-lo amb una sèrie de dades. D'aquesta manera es pot crear un model per a aquest parlant, és a dir, una mena d'ADN (tot i que molt menys exacte) que l'identificarà posteriorment. Si després tenim les dades d'un altre locutor i volem saber si és el mateix que hem enregistrat, només caldrà que comparem els models de cadascun dels locutors.

Per a aquests experiments en concret, cada model de locutor ha estat entrenat amb vuit converses. [3]En total, s'han obtingut mesures amb 1.343 clients i 2.381 impostors; és a dir, un total de 3.724 valors.

4.3. Característiques prosòdiques

Els experiments de verificació desenvolupats sobre les característiques prosòdiques ens permeten veure quin és el grau de fiabilitat de cadascuna de les característiques. La taula 1 mostra els valors dels EER obtinguts per a les vuit primeres característiques prosòdiques descrites a l'apartat 3. En aquest cas, les pauses no s'han tingut en compte perquè, en tractar-se de converses entre parelles de locutors, poden dependre del locutor que no s'està analitzant.

Taula 1. EER per a la fusió de característiques prosòdiques i espectrals

Característica	EER (%)
log (trames/paraula)	30,4
trames sonores	31,6
log F0_mitjana	19,2
log F0_màxima	21,4
log F0_mínima	21,5
log rang F0	27,2
pendent F0	39,1
pendent estilitzat F0	28,7
fusió	**14,9**

Com podem observar, les característiques que funcionen millor com a identificadors biomètrics —és a dir, les que tenen un EER més baix— són les relacionades amb els valors mitjà, màxim i mínim de la freqüència fonamental.

Addicionalment, la combinació de totes les característiques mitjançant el matcher weighting i normalitzant prèviament amb la tècnica min-max, ens proporciona un EER del 14,9 %.

[3] D'una o més converses es poden extreure les dades necessàries per crear un model, calculant la F0 mitjana, etc. En aquest experiment se n'han utilitzat vuit.

4.4. Combinació del sistema prosòdic amb un sistema espectral

Com ja s'ha comentat anteriorment, la forma espectral del senyal de veu conté informació sobre el tracte vocal i la font d'excitació de la glotis mitjançant els formants i la freqüència fonamental, respectivament. Per aquest motiu, la majoria de paràmetres utilitzats en el RAL s'obtenen de l'espectre del senyal.

En aquest apartat es fa servir un sistema de verificació basat en coeficients espectrals per tal de veure en quin grau la utilització dels elements prosòdics el poden complementar per ajudar a millorar-ne el funcionament. Aquest sistema espectral consisteix en un model de mescla de Gaussianes de 32 components, que utilitza vectors que consten de 20 paràmetres de *Filtratge Freqüencial* (FF) (Nadeu et al. 1995), amb trames de 30 milisegons i un desplaçament de 10 milisegons. Seguint el mateix protocol que en els paràmetres prosòdics, amb aquest sistema s'obté un EER del 10,1 % en els experiments de verificació.

Normalitzant les mesures espectrals i les prosòdiques amb la tècnica min-max i fusionant-les mitjançant el matcher weighting, s'obtenen els resultats següents:

Taula 2. EER per a la fusió de característiques prosòdiques i espectrals

característiques	EER (%)
prosòdiques	14,9
espectrals	10,1
fusió	**7,7**

Així doncs, els resultats fan evident la complementarietat dels dos tipus de paràmetres, ja que el fet de combinar-los millora considerablement el resultat final del sistema.

5. CONCLUSIONS

Tot i que la biometria és una ciència molt antiga —diuen que al segle v aC ja s'utilitzaven impressions en argila de les

empremtes dactilars a Babilònia per fer operacions de negoci—, no ha estat fins als últims anys que el fet d'explotar-la en l'àmbit comercial ha esdevingut considerable. Un dels objectius d'aquest article ha estat il·lustrar, de la manera més planera possible, les característiques principals d'aquesta metodologia i les seves aplicacions.

Però l'article també ha pretès esmentar la necessitat d'utilitzar el coneixement lingüístic en mètodes que, com la biometria, fins ara havien quedat relegats a l'àmbit de l'enginyeria. Sense voler entrar en detalls, s'ha volgut il·lustrar com es pot millorar un sistema de reconeixement del locutor —que no és més que un tipus específic de sistema de reconeixement biomètric—, mitjançant la incorporació de característiques prosòdiques, és a dir, d'una part del coneixement lingüístic que els éssers humans utilitzem per a reconèixer-nos els uns als altres.

Si bé pot semblar que els experiments descrits s'han limitat a fer servir la veu com a identificador biomètric i que no s'ha il·lustrat de quina manera la prosòdia pot millorar sistemes biomètrics basats en altres identificadors, sí que és cert que en altres estudis, com el de Farrús (2008), s'ha utilitzat la prosòdia en altres tipus de sistemes biomètrics —en aquest cas la cara—, on es demostra que la prosòdia és també molt útil per millorar-los.

En definitiva, les pretensions d'aquest article es limiten a donar a conèixer la importància i la necessitat de la lingüística en l'àmbit de la biometria i, específicament, del reconeixement del locutor.

Agraïments

L'autora vol agrair a Ramon Cerdà el seu suport i les seves valuosíssimes aportacions a l'hora de redactar aquest article.

6. Referències bibliogràfiques

Adami, A. (2007). «Modeling prosodic differences for speaker recognition». *Speech Communication,* 49, 4, 277-291.

Andrews, W., Kohler, M.A., Campbell, J.; Godfrey, J. i Hernández-Cordero, J. (2002). «Gender-dependent phonetic refraction for speaker recognition». *Proceedings of the ICASSP*, 1, 149-152.

Atkinson, J.E. (1978). «Correlation analysis of the physiological factors controlling fundamental voice frequency». *Journal of the Acoustical Society of America*, 63, 1, 211-222.

Bartkova, K., Le-Gac, D., Charlet, D. i Jouvet, D. (2002). «Prosodic parameter for speaker identification». *Proceedings of the ICSLP*, 1197-1200.

Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K. i Senior, A.W. (2004). *Guide to Biometrics.* New York: Springer.

Brand, J.D., Mason, J.S. i Colomb, S. (2001). «Visual speech: a physiological or behavioural biometric?». *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication. Lecture Notes in Computer Science*, 2091, 157-168.

Cicres, J. (2007). «Análisis discriminante de un conjunto de parámetros fonético-acústicos de las pausas llenas para identificar hablantes». *Síntesis Tecnológica.* Universidad Austral de Chile, 3, 2, 87-96.

Doddington, G. (2001). «Speaker recognition based on idiolectal differences between speakers». *Proceedings of the Eurospeech*, 4, 2521-2524.

Farrús, M. (2008). *Fusing prosodic and acoustic information for speaker recognition.* Tesi doctoral. Barcelona; Universitat Politècnica de Catalunya. Departament de Teoria del Senyal i Comunicacions.

Godfrey, J.J., Holliman, E.C. i McDaniel, J. (1990). «Switchboard: Telephone speech corpus for research and development». *Proceedings of the ICASSP.* Alburquerque: NM, 1, 517-520.

HERNANDO, J. (2001). «Reconocimiento automático de locutores». *Ciencia & Tecnología del siglo XXI*. Barcelona: Tibidabo, 2, 23-36.

INDOVINA, M., ULUDAG, U., SNELIK, R., MINK, A. i JAIN, A. (2003). «Multimodal biometric authentication methods: A COTS approach». *Proceedings of the Workshop on Multimodal User Authentication*. Santa Barbara: CA, 99-106.

JAIN, A.K., ROSS, A. i PRABHAKAR, S. (1994). «An introduction to biometric recognition». *IEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-based Biometrics*, 14, 1, 4-20.

MALTONI, D., MAIO, D., JAIN, A.K. i PRABHAKAR, S. (2003). *Handbook of Fingerprint Recognition.* New York: Springer.

NADEU, C., HERNANDO, J. i GORRICHO, M. (1995). «On the decorrelation of filter bank energies in speech recognition». *Proceedings of the Eurospeech*. Madrid, 1381-1384.

PESKIN, B., NAVRÁTIL, J., ABRAMSON, J., JONES, D., KLUSÁČEK, D., REYNOLDS, D.A. i XIANG, B. (2003). «Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02». *Proceedings of the ICASSP*. Hong Kong, China, 4, 792-795.

REYNOLDS, D.A., ANDREWS, W., CAMPBELL, J., NAVRÁTIL, J., PESKIN, B., ADAMI, A., JIN, Q., KLUSÁČEK, D., ABRAMSON, J., MIHAESCU, R., GODFREY, J., JONES, D. i XIANG, B. (2003). «The SuperSID project: exploiting high-level information for high-accuracy speaker recognition». *Proceedings of the ICASSP*. Hong Kong, China, 4, 784-787.

SCHMIDT-NIELSEN, A. i CRYSTAL, T.H. (2000). «Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 Speaker Evaluation Data». *Digital Signal Processing*, 10, 1-3, 249-266.

SHRIBERG, E., STOLCKE, A., HAKKANI-TUR, D. i TUR, G. (2000). «Prosody-based automatic segmentation of speech into sentences and topics». *Speech Communication*, 32, 1-2, 127-154.

WEBER, F., MANGANARO, L., PESKIN, B. i SHRIBERG, E. (2002). «Using prosodic and lexical information for speaker identification». *Proceedings of the ICASSP,* Orlando, 1, 141-144.

WERTZNER, H.F., SCHREIBER, S. i AMARO, L. (2005). «Analysis of the fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders». *Brazilian Journal of Otorhinolaryngology*, 71, 5, 582-588.