Environnement de développement MATLAB R2013a

Le domaine de la reconnaissance automatique de la parole est actuellement très actif. De nombreux laboratoires de recherche et des industriels effectuent des recherches dans ce domaine, avec un souci théorique et applicatif très marqué. Même si quelques problèmes de reconnaissance comme la reconnaissance de mots isolés avec un vocabulaire limité et prononcés dans des conditions calmes d'utilisation ou la reconnaissance dépendant du locuteur peuvent être considérés comme ayant atteint un niveau de performance satisfaisant, la reconnaissance automatique mérite encore de nombreux travaux de recherche pour étendre son champ d'application. Un axe important de recherche concerne l'amélioration de la robustesse d'un système de reconnaissance lorsque l'environnement de test est sensiblement différent de l'environnement d'apprentissage. Ce sujet a été le centre d'attention de ce document. Deux aspects du problème de robustesse ont été présentés : la robustesse au bruit et la robustesse au locuteur.

Nos travaux de recherche ont porté sur la fusion d'informations acoustiques et visuelles pour la RAP. Nous avons donc abordé les principaux problèmes sous-jacents à cette fusion, à savoir la paramétrisation des informations de parole et la nature des systèmes de reconnaissance dans chacune des modalités, ainsi que le lieu et la nature du processus de fusion des informations sensorielles. Nous avons choisi de résoudre ces problèmes en nous appuyant sur des études réalisées dans le domaine de la perception audiovisuelle de la parole.
Nous avons développé différents systèmes pour effectuer la fusion des informations acoustiques et visuelles en prenant appui sur des modèles perceptifs. Ces systèmes ont été testés sur deux corpus audiovisuelles CUAVE.

7.2 Perspectives

Les travaux commencés au cours de cette thèse ouvrent la voie à de nombreux travaux futurs.

- La prise en compte de la parole continue ainsi spontanée est vitale pour un système de reconnaissance grand public.
- Les pauses, les répétitions, les hésitations, les phrases en suspens posent des problèmes par la suite aux autres modules de l'application visée.

- Les gens utiliseront les systèmes de reconnaissance à condition que le taux d'erreur de reconnaissance soit suffisamment faible. La reconnaissance robuste est donc nécessaire. L'utilisation d'un système de reconnaissance dans un milieu bruité et par différentes personnes devrait être habituelle.
- La prise en compte des bruits non stationnaires, dont l'importance a été soulevée à travers ce document, nécessite de continuer l'effort engagé. Nous n'en sommes qu'au début. L'étude des problèmes de détections de changement des bruits et la prise en compte de ces moments pendant la reconnaissance doit se poursuivre.
- Avec la représentation par adjacence, présentée dans le 4ème chapitre, nous avons établi que le manque de compatibilité entre le GA d'une part et l'opérateur de mutation génétique défini sur la base d'approches déterministes d'autre part, nuisait à l'efficacité de l'approche. C'est donc prioritairement sur ce point que devront se focaliser de futurs développements.

Annexe A

Environnement de développement: MATLAB R2013a

MATLAB (« matrix laboratory ») est un langage de programmation de quatrième génération émulé par un environnement de développement du même nom ; il est utilisé à des fins de calcul numérique. Développé par la société américaine The MathWorks, MATLAB permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes, de créer des interfaces utilisateurs, et peut s'interfacer avec d'autres langages comme le C, C++, Java, et Fortran. Les utilisateurs de MATLAB (environ un million en 20041) sont de milieux très différents comme l'ingénierie, les sciences et l'économie dans un contexte aussi bien industriel que pour la recherche. Matlab peut s'utiliser seul ou bien avec des toolbox (« boîte à outils »).

Le logiciel Matlab® et l'environnement graphique interactif Simulink® sont particulièrement performants et adaptés à la résolution de problèmes d'automatique, notamment pour la modélisation et la simulation des systèmes dynamiques.

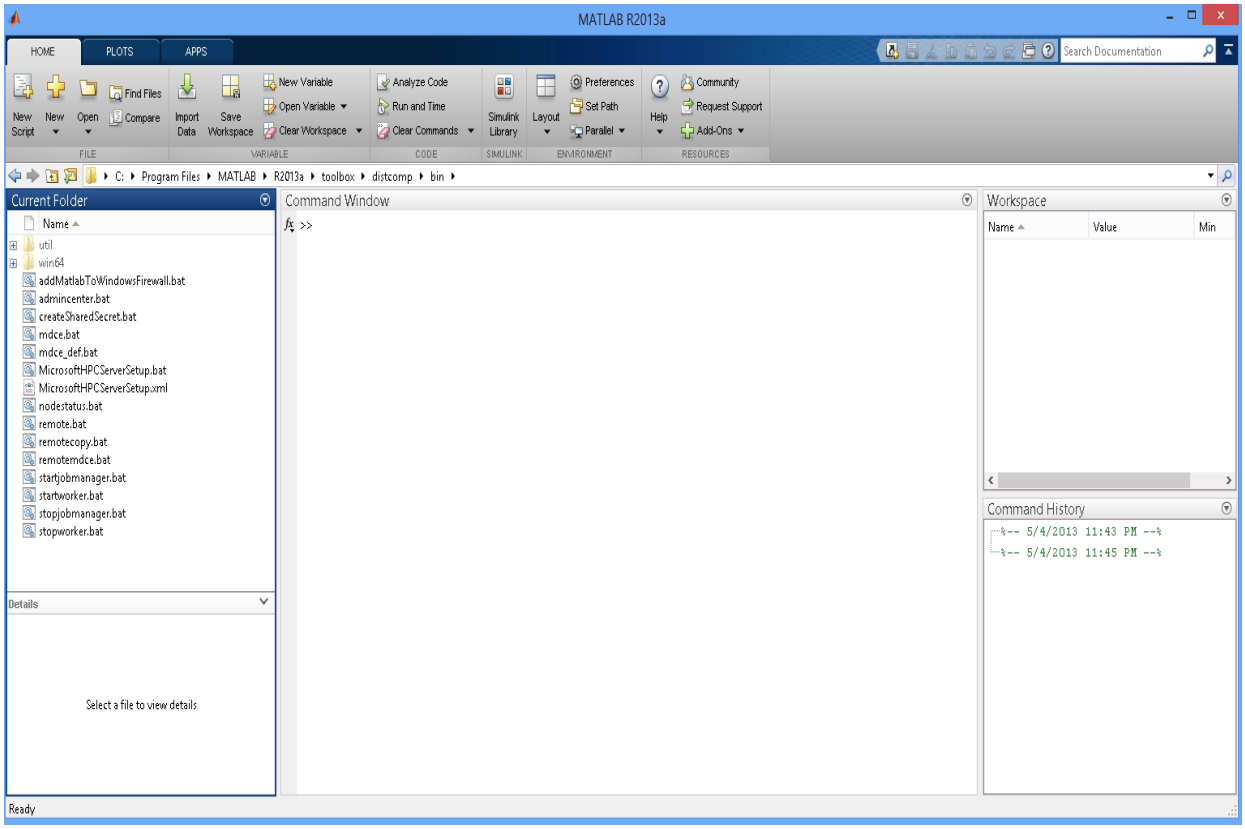

Figure A.1 – *L'interface de l'environnement Matlab (R2013a).*

- **Avantages** :
  - collection très riche de librairies avec de nombreux algorithmes, dans des domaines très variés. Exécution rapide car les librairies sont souvent écrites dans un langage compilé.
  - environnement de développement très agréable : aide complète et bien organisée, éditeur intégré, etc.
  - support commercial disponible

- **Inconvénients :**
  - langage de base assez pauvre, qui peut se révéler limitant pour des utilisations avancées.
  - prix élevé

- **Pourquoi alors Matlab ?**

En effet plusieurs extensions plus « pointues » ont été conçues sous la forme de « TOOLBOXes », qui sont des paquets (payants) de fonctions supplémentaires dédiées à des domaines aussi variés que les statistiques, le traitement du signale et d'image, la logique floue, les réseaux de neurones, les ondelettes,… et qui permettent de résoudre un bon nombre de problèmes relatifs à ses domaines. Pour visualiser ces fonctions, il suffit de taper **help** suivi du nom de la famille à laquelle appartient la fonction. Pour connaitre le nom de ces familles, il suffit juste de taper **help**. Il comporte plus de 1500 fonctions préprogrammées.

- **bibliothèques utilisés :**

La phase d'apprentissage est réalisée en deux étapes majeures : l'initialisation et la ré-estimation. Nous les avons conçus à partir de la plateforme HTK (Hidden Markov Model ToolKit) de l'Université de Cambridge. La boîte à outils HTK est efficace, flexible (liberté du choix des options et possibilité d'ajout d'autres modules) et complète dans le sens ou elle fournit une documentation très détaillée (le livre HTK (Young et al. 2006) est une encyclopédie dans le domaine).

Structure et fonctionnement du logiciel

Ce logiciel traite une phase importante de tout type de reconnaissance de formes qui est la phase de reconnaissance. Il implémente précisément deux méthodes de prétraitement (DCT et RASTA-PLP) et l'algorithme K-means pour le clustering, ainsi 2 méthodes de reconnaissance HMM et le modèle hybride GA/HMM.

Le logiciel est implémenté sur Matlab R2013a, il est sous formes de fichier script MATLAB, ces fichiers MATLAB qui ont l'extension (.m) peuvent être considérés comme des fonctions qui peuvent être appelé à partie de l'interpréteur de commande MATALAB et qui se servent à leur tour d'un autre type de fichier des fichiers qui ont l'extension (.mat). Ces derniers fichiers représentent dans MTLAB des bases de données.

Notre application contient un fichier principale qui fait appelle aux autre fichiers .Ce fichier est nommé "interface " (voir figure A.2).

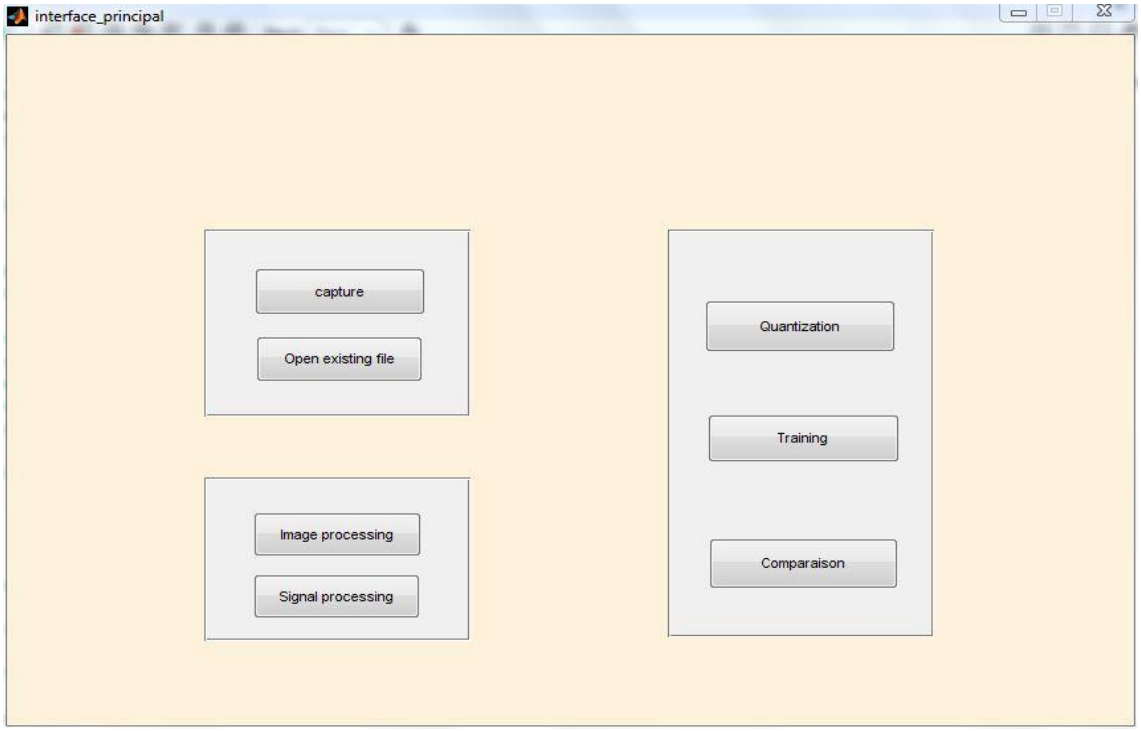

Figure A.2 – Interface principale du logiciel.

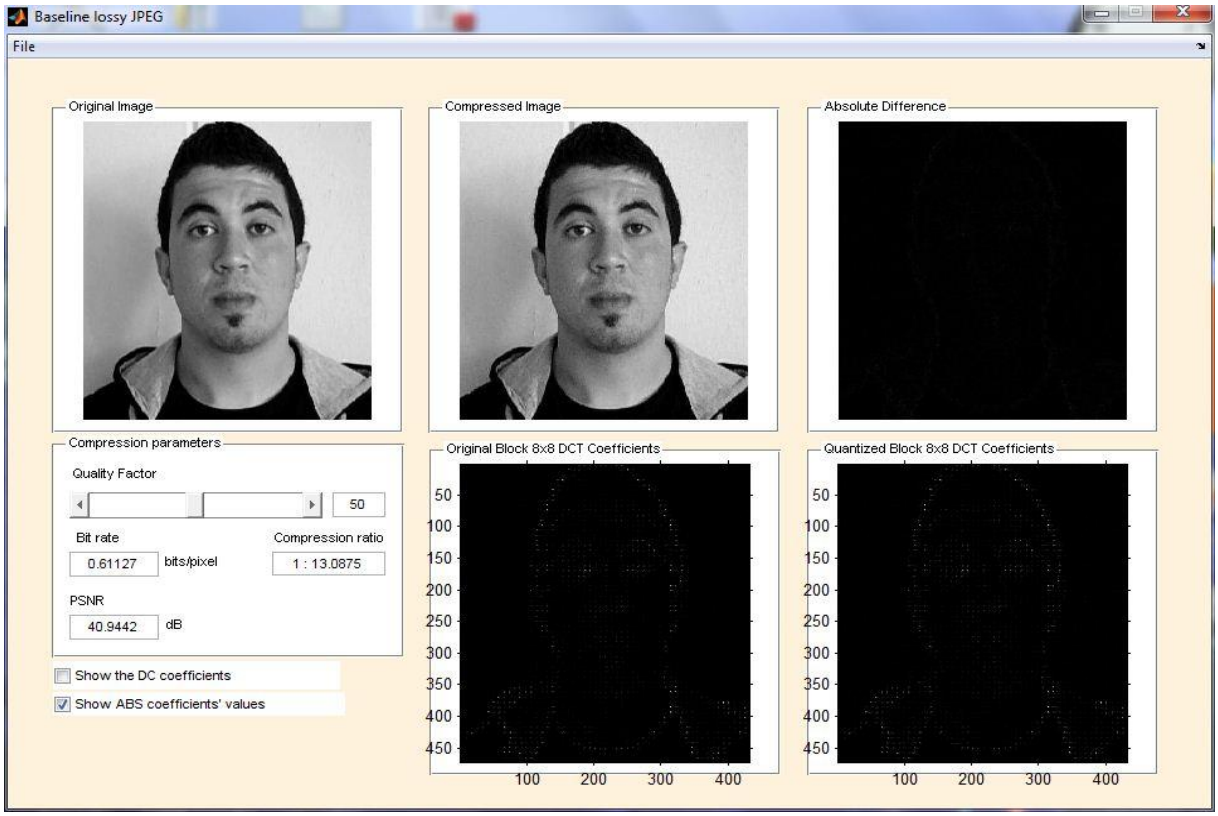

Figure A.3 – Interface d'extraction des paramètres visuels.

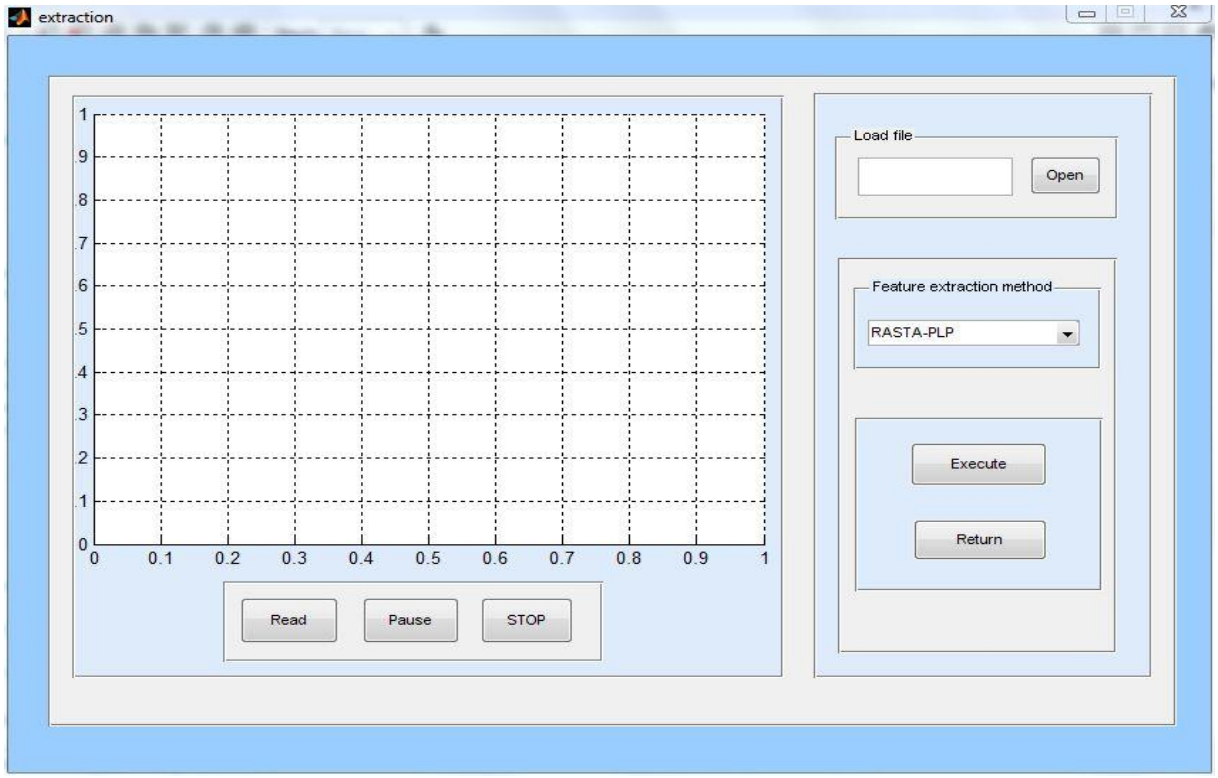

Figure A.4 – Interface d'extraction des paramètres acoustiques.

Bibliographie

Abry C., Boë L.-J., Corsi P., Descout R., Gentil M. and Graillot P. (1980). Labialité et Phonétique, publications de l'Université des langues et lettre de Grenoble.

Adjoudani, A., Guiard-Marigny, T., Le Goff, B. and Benoît, C. (1994). Un modèle 3d de lèvres parlantes. *In Actes des XX° Journées d'Etude sur la Parole (JEP)*, pp. 143–146.

Adjoudani, A. and Benoît, C. (1995). Audio-visual speech recognition compared across two architectures, *in Proc. of the 4th EUROSPEECH Conference*, Madrid, Espagne, pp. 1563-1566.

Adjoudani, A. (1998). Reconnaissance automatique de la parole audiovisuelle. *Thèse de doctorat*, Institut National Polytechnique de Grenoble.

Allegre, J. (2003). Approche de la reconnaissance automatique de la parole. *Rapport cycle probatoire, CNAM.*

Alpaydin, E. (2004). Introduction to machine learning. *MIT Press*.

Basso, A. Graf, H.P., Gibbon, D., Cosatto, E. and Liu, S. (2001). Virtual light: Digitally-generated lighting for video conferencing applications. In Proc. ICIP, 2: pp. 1085-1088, Thessaloniki, Greece, October 7-10.

Benoît, C., Guiard-Marigny, T., Le Goff, B. and Adjoudani, A. (1996). Which Components of the Face Do Humans and Machines Best Speechread?, in Speechreading *by Humans and Machines, D. Stork and M. Hennecke* (eds.), Springer-Verlag, Berlin, pp. 351-372.

Binnie C.A., Montgomery A.A. and Jackson P.L. (1974). Auditory and visual contributions to the perception of consonants, *Journal of Speech & Hearing Research*, 17, pp. 619-630.

Berger, K. W., Garner, M., and Sudman, J. (1971) . The effect of degree of facial exposure and the vertical angle of vision on speechreading performance. *Teacher of the Deaf*, 69: pp. 322–326.

Beyer, H.-G. (2001). The Theory of Evolution Strategies. *Natural Computing Series*. Springer, Heidelberg.

Bregler, C., Hild, H., Manke, S. and Waibel, A. (1993). Improving connected letter recognition by lipreading, *Proc of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, IEEE*, 1, pp. 557-560.

Bridges, C.L. and Goldberg, D.E. 1991. An analysis of multipoint crossover. In *Proceedings of the Foundation Of Genetic Algorithms*. FOGA.

Bogert, B., Healy, M. and Tukey, J. (1963). The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. Time Series Analysis, pp. 209-243.

Boite, R., Bourlard, H., Dutoit, T., Hancq, J. and Leich, H. (2000). *Traitement de la parole* (Presses Polytechniques et Universitaires Romandes, Lausanne).

Bouchet, A. and Cuilleret, J. (1972). Anatomie topographique descriptive et fonctionelle, Villeurbanne, Simep éditions.

Broun, C.C., Zhang, X., Mersereau, R.M. and Clements, M. (2002). Automatic speechreading with application to speaker verification. *In Proc. ICASSP*, 1: pp. 685-688, Orlando, FL, USA, May 13-17.

Brunelli, R. and Poggio, T. (1993). Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042-1052.

Burnham, D. and Dodd, B. (1996). Auditory-visual speech perception as a direct process: the McGurk effect in infants and across languages, Speechreading by Humans and Machines, Stork et

Hennecke (eds.), Springer-Verlag, Berlin, pp. 103-114.

Cathiard, M.A. (1988).Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français. Mémoire de maîtrise, Université Grenoble II.

Cathiard, M.A. (1989). La perception visuelle de la parole : aperçu des connaissances, Bulletin de l'Institut de Phonétique de Grenoble, 18: pp. 109-193.

Cathiard, M.A. (1994). La perception visuelle de l'anticipation des gestes vocaliques : cohérence des événements audibles et visibles dans le flux de la parole. Thèse de doctorat de psychologie cognitive, UFR SHS, Université Pierre Mendès France.

Chan, M.T., Zhang, Y. and Huang, T.S. (1998). Real-time lip tracking and bimodal continuous speech recognition. In Proc. 2nd MMSP, pp. 65-70, Los Angeles, CA, USA, December 7-9.

Chiou, G.I. and Hwang, J.-N. (1996). Lipreading from color motion video. *In Proc. ICASSP*, 4: pp. 2158-2161, Atlanta, GA, USA.

Coianiz, T., Torresani, L. and Caprile, B. (1996). 2D deformable models for visual speech analysis. In Stork and Hennecke (1996), pp. 391-398.

Collen, P., Rault, J.B. and Betser, M. (2007). Phase estimating method for a digital signal sinusoidal simulation," Software Patent PCT/FR2006/051361, 2007.

Dai, Y. and Nakano, Y. (1996). Face-Texture Model Based on SGLD and Its Application in Face Detection in a Color Scene. *Pattern Recognition* 29(6), pp. 1007-1017.

Dallos, P. (1973). The Auditory Periphery: Biophysics and Physiology. *New York, USA: Academic Press.*

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. *Londres, John Murray*.

Davis, S. and Melmerstein, P. (1980). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on ASSP*, 28: pp. 357-366.

Demuynck, K., Garcia, O. and Van Compernolle, D. (2004). Synthesizing speech from speech recognition parameters. *Proc. of ICSLP*.

Deviren, M. (2004). Systèmes de reconnaissance de la parole revisités : Réseaux Bayésiens dynamiques et nouveaux paradigmes. *Université de Nancy, Nancy, Thèse de doctorat*.

Dodd, B. and Campbell, R. (1987) (eds.), Hearing by Eye: The Psychology of Lipreading, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Duchnowski, P., Hunke, M. Büsching, D., Meier, U. and Waibel, A. (1995). Toward movement-invariant automatic lip-reading and speech recognition. In Proc. ICASSP, 1: pp.109–112, Detroit, MI, USA.

Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141-151.

Erber N.P. (1974). Effect of angle, distance, and illumination on visual reception of speech by profoundly deaf children. *Journal of Speech and Hearing Research*, 17:pp. 99–112.

Erber N.P. (1975). Auditory-visual perception of speech, *Journal of Speech and Hearing Disorders*, 40, pp. 481-492.

Escudier, P., Benoît, C. and Lallouache, M.T. (1990). Identification visuelle de stimuli associés à l'opposition /i/ - /y/: étude statistique, Proceedings of the First French Conference on Acoustics, Lyon, France, pp. 541-544.

Eyben, F., Wöllmer, M. and Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proc. of ACM Multimedia, pp. 1459-1462.

Fant, G. (1973). Speech Sounds and Features », M.I.T. Press, Cambridge, USA.

Fogel, L.J., Owens, A.J. and Walsh, M.J. (1966). Artificial Intelligence through Simulated Evolution. *Wiley*, New York.

Goh, J., Tang, L. and Al turk, L. (2010). Evolving the Structure of Hidden Markov Models for Micro aneurysms Detection. *UK Workshop on Computational Intelligence (UKCI)*, pp.1–6.

Goldberg, D. and Richardson, J. (1987). Genetic algorithm with shearing for multi-model function optimization, *In J.J. Proceeding of the 2nd international conference on genetic algorithms*, pp. 41-49, Lawrence Erlbaum associates.

Goldberg, D. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. *Addison Wesley Reading, Massachusetts*.

Goldberg, D. (1991). Real-coded genetic algorithms, virtual alphabets and blocking. Complex Systems, 5: pp. 139-167.

Gouet, V. and Montesinos, P. (2002). Normalisation des images en couleur face aux changements d'illumination. *In Proc. RFIA'02*, 2: pp. 415-424, Angers, France, January 8-10.

Gray, M.S., Movellan, J.R. and Sejnowski, T.J. (1997a). A comparison of local versus global image decompositions for visual speechreading. *In Proc. 4th Annual Joint Symposium on Neural Computation*, pp. 92-98, Pasadena, CA, USA, May 17.

Gray, M.S., Movellan, J.R. and Sejnowski, T.J. (1997b). Dynamic features for visual speechreading: A systematic comparison. *In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, ANIPS*, 9: pp. 751-757. The MIT Press.

Gupta, M. and Garg, Dr.A.K. (2012). Analysis of image compression algorithm Using DCT. *International Journal of Engineering Research and Applications (IJERA)*, 2(1): pp.515–521.

Gurbuz, S., Patterson, E.K., Tufekci, Z. and Gowdy, J.N. (2001a). Lip-reading from parametric lip contours for audio-visual speech recognition. *In Proc. 7th Eurospeech*, 2: pp.1181-1184, Aalborg, Denmark, September 3-7.

Gurbuz, S., Patterson, E.K., Tufekci, Z. and Gowdy, J.N. (2001b). Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition. In Proc. ICASSP, 1: p. 177-180, Salt Lake City, UT, USA, May 7-11.

Hlaoui, A. (1999). Reconnaissance de mots isolés arabes par hybridation de réseaux de neurones et modèles de Markov cachés. *École nationale d'ingénieurs de Tunis*.

Hardcastle, W.J. (1976). Physiology of Speech Production, Academic Press, Londres.

Harvey, R., Matthews, L., Bangham, J.A. and Cox, S. (1997). Lip reading from scale-space measurements. In Proc. CVPR, pp. 582-587, Puerto Rico, June.

Haton, J.-P. (2006). Reconnaissance automatique de la parole : Du signal à son interprétation. *Dunod Paris*.

Hermansky, H., Morgan, N., Bayya, A. and Kohn, P. (1992). RASTA-PLP Speech Analysis. *IEEE International conference on Acoustics, speech and signal processing*, 1: pp.121–124.

Holland, J. (1975). Adaptation in Natural and Artificial Systems. *University of Michigan Press*.

Hunke, H. M. and Waibel, A. (1994). Face locating and tracking for human-computer interaction, *Proc. Twenty-Eight Asilomar Conference on Signals, Systems & Computers*, Monterey, CA, USA.

Hunke, H. M. (1994). Locating and tracking of human faces with neural networks. Master's thesis, University of Karlsruhe.

Jacob, B. and Sénac, C. (1996). Un modèle maître-esclave pour la fusion de données acoustiques et articulatoires en reconnaissance. *In Actes des Journées d'Etude sur la Parole (JEP)*, pp. 363–366, Avignon, Juin.

Jakiela, M., Chapman, C., Duda, J., Adweuya, A. and Saitou, K. (2000).Continuum structural topology design with genetic algorithm. *Comput. Methods Appl. Mech. Engrg* 186, pp. 339-356.

Jourlin, P. (1996). Handling disynchronization phenomena with hmm in connected speech. *In Proceedings of European Signal Processing Conference*, pp. 133–136, Trieste.

Kant, E. (1787). Critique de la Raison Pure, *Presses Universitaires de France, 11ème edition*, 1944, édition originale, 1787.

Khandait, S.P., Khandait, P.D. and Thool, Dr.R.C. (2009). An Efficient Approach to Facial Feature Detection for Expression Recognition. *International Journal of Recent Trends in Engineering*, 2(1): pp.179–**182.**

Kicinger, R., Arciszewski, T., and Jong, K. D. (2005). Evolutionary computation and structural design: A survey of the state-of-the-art. Computers & Structures, 83(23-24): pp. 1943-1978.

Klatt. D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal Phonetique.* 7: pp. 279–312.

Kubrick, S. (1968). 2001 : A space odyssey (2001 : l'odysée de l'espace). Metro-Goldwyn-Mayer (Turner Entertainment Co), April 3. http ://www.kubrick2001.com/ , http://sfstory.free.fr/films/2001.html.

Kuhl, P.K. and Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. Science, 218, pp. 1138-1141.

Kwong, S. and Chau, C.W. (1997). Analysis of Parallel Genetic Algorithms on HMM Based Speech Recognition System. *IEEE Transactions on Consumer Electronics.* 43(4): pp. 1229 – 1233.

Ladefoged P. (1979). Articulatory parameters, W.P.P. 45, U.C.L.A., pp. 25-31.

Lallouache M.T. (1991). Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres, PhD. dissertation, INPG, Grenoble, France.

Laprie, Y. (2000). Analyse spectrale de la parole.

Larr A. L. (1959). Speechreading through closed-circuit television. *Volta Review,* 61: pp.19–21.

Lee, J. and Kim, J.Y. (2001). An efficient lipreading method using the symmetry of lip. *In Proc. 7th Eurospeech*, 2: pp. 1019-1022, Aalborg, Denmark, September 3-7.

Le Goff, B., Guiard-Marigny, T., and Benoît, C. (1995). Read my lips ... and my jaw! how intelligible are the components of a speaker's face ? In Eurospeech'95, Madrid, Spain.

Le Goff, B., Guiard-Marigny, T., and Benoît, C. (1996).Progress in Speech Synthesis, *chapitre Analysis-synthesis and intelligibility of a talking face*, pp. 235–246. Springer, New York.

Le Huche, F. and Allali, A. (2001). *La Voix. Anatomie et physiologie des organes de la voix et de la parole* (Masson, Paris).

Leroy, B. and Herlin, I.L. (1995). Un modèle déformable paramétrique pour la reconnaissance de visages et le suivi du mouvement des lèvres. *In 15th GRETSI Symposium Signal and Image Processing*, pp. 701-704, Juan-les-Pins, France, September 18-21.

Leroy, B. Chouakria, A., Herlin, I.L. and Diday, E. (1996a). Approche géométrique et classification pour la reconnaissance de visages. In Proc. RFIA, pp. ??-??, Rennes, France.

Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech production revised. *Cognition*, 21: pp.1–36, 1985.

Lievin, M. and Luthon, F. (1999). Lip features automatic extraction. Proceedings of IEEE International Conference on Image Processing, Chicago, IL, USA, 3: pp. 168–172.

Liew, A.W.C., Sum, K. L., Leung, S.H. and Lau, W.H. (1999). Fuzzy segmentation of lip image using cluster analysis. In Proc. 6th Eurospeech, 1: pp. 335-338, Budapest, Hungary, September 6-9.

Liu, L., He, J. and Palm, G. (1997). Effects of the phase on the perception of intervocalic stop consonants. Speech Communication, 4(22): pp. 403-417.

Lockwood, P., Boudy, J. and Blanchet, M. (1992). Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments. *Proc. of IEEE ICASSP*, 1: pp. 265-268.

Luettin, J. Thacker, N.A. and Beet, S. (1996a). Active shape models for visual speech feature extraction. *In Stork and Hennecke* (1996), pp. 383-390.

Luettin, J. Thacker, N.A. and Beet, S. (1996b). Locating and tracking facial speech features. *In Proc. ICPR*, 1: pp. 652-656, Vienna, Austria, August 25-29.

Luettin, J. Thacker, N.A. and Beet, S. (1996c). Speaker identification by lipreading. *In Proc. 4th ICSLP*, 1: pp. 62-65, Philadelphia, PA, USA, October 3-6.

Luettin, J. Thacker, N.A. and Beet, S. (1996d). Speechreading using shape and intensity information. In Proc. 4th ICSLP, 1: pp. 58-61, Philadelphia, PA, USA, October 3-6.

Luettin, J. Thacker, N.A. and Beet, S. (1996e). Statistical lip modelling for visual speech recognition. *In Proc. 8th Eusipco*, 1: pp. 137-140, Trieste, Italy, September 10-13.

Luettin, J. Thacker, N.A. and Beet, S. (1996f). Visual speech recognition using active shape models and hidden Markov models. *In Proc. ICASSP*, 2: pp. 817-820, Atlanta, GA, USA, May 7-10.

Luettin, J. and Thacker, N.A. (1997). Speechreading using probabilistic models. Computer Vision and Image Understanding, 65(2):163-178.

Luettin, J. (1997a). Towards speaker independent continuous speechreading. *In Proc. 5th Eurospeech*, pp. 1991-1994, Rhodes, Greece, September 22-25.

Luettin, J. (1997b). Visual Speech and Speaker Recognition, *PhD dissertation*, Université de Sheffield.

Luettin, J. and Dupont, S. (1998). Continuous audio-visual speech recognition. *LNCS*, 1407: pp. 657-673.

Makhlouf A., Lazli, L. and Bensaker, B. (2013a). Automatic Speechreading Using Genetic Hybridization of Hidden Markov Models. *In Proceeding of the IEEE World Congress on Computer and Information Technology (WCCIT'13)*, June 22-24, 2013, Sousse, Tunisia.

Makhlouf A., Lazli, L. and Bensaker, B. (2013b). Hybrid Hidden Markov Models and genetic algorithm for Robust Automatic visual speech recognition. *Journal of Information Technology Review (JITR)*, 4(3): pp. 105-114.

Makhlouf A., Lazli, L. and Bensaker, B. (2016). Structure Evolution of Hidden Markov Models for Audiovisual Arabic Speech Recognition. *International Journal of Signal and Imaging Systems Engineering, IJSISE, 9(1)*.

Malasné, N., Yang, F., Paindavoine, M.and Mitéran, J. (2002). Suivi dynamique et vérification de visages en temps réel : algorithme et architecture. *In Proc. RFIA'02*, pp.77-86, Angers, France.

Mase, K. (1991). Automatic lipreading by optical-flow analysis. Systems and Computers in Japan, 22(6): 67-75.

Massaro, D.W. (1987). Categorical Perception: The Groundwork of Cognition, chapitre Categorical partition: a fuzzy logical model of categorization behavior. *Cambridge, MA : University Press.*

Massaro, D.W. (1989). Multiple book review of Speech perception by ear and eye, Behavioral and Brain Sciences, 12, pp.741-794.

Massaro, D.W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. *Cambridge, Massachusetts : MIT Press*.

Matthews, L. Bangham, J. and Cox, S. (1996a). Audiovisual speech recognition using multiscale nonlinear image decomposition. In Proc. 4th ICSLP, 1: pp. 38-41, Philadelphia, PA, USA, October 3-6.

Matthews, L. Bangham, J.A., Harvey, R. and Cox, S. (1998). A comparison of active shape models and scale decomposition based features for visual speech recognition. *LNCS*, 1407: pp. 514-528.

McGurk, H. and McDonald, J. (1976). Hearing Lips and Seeing Voices, Nature, 264: pp. 746-748.

Meier, U. Hürst, H. and Duchnowski, P. (1996). Adaptive bimodal sensor fusion for automatic speechreading. *In Proc. ICASSP*, pp. 833-836, Atlanta, GA, USA, May.

Messer, k., Matas, J., Kittler, J., Luettin, J. and Maître, G. (1999). XM2VTSDB : The extended M2VTS database. *In Proc. 2nd AVBPA*, pp. 72–77, Washington, DC, USA, March 22-23.

Michalewicz, Z. and Janikov, C.Z. (1991). Handling constraints in genetic algorithms. In *Proceedings of the Fourth International Conference on Genetic Algorithm*. ICGA.

Milner, B. and Darch, J. (2011). Robust Acoustic Speech Feature Prediction From Noisy Mel-Frequency Cepstral Coefficients. IEEE Trans. on ASLP, 2(19): pp. 338-347.

Movellan, J.R (1995). Visual speech recognition with stochastic networks. *In Gerald Tesauro, David Touretzky, and Todd Leen, editors, ANIPS*, 7: pp. 851-858, Cambridge, MA, USA. *The MIT Press*.

Movellan, J.R and Chadderdon, G. (1996). Speechreading by Man and Machine: Models, Systems and Applications. *chapitre Channel separability in the audiovisual integration of speech : A Bayesian approach*, pp. 473–488. *Springer-Verlag, NATO ASI Series, Berlin, Germany.*

Murty, K.S.R. and Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters, 1(13): pp. 52-55.

Nakano, Y. (1961). A study on the factors which influence lipreading of deaf children. *Language research in countries other than the United States, Volta Review*, 68:pp. 68–83. Cited by Quigley (1966).

Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *Journal of Acoustic Society of America*, 28: pp.1275–1277.

Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C. and Murphy, K. (2002). A coupled HMM for audio-visual speech recognition. *In Proc. ICASSP*, 2: pp. 2013-2016, Orlando, FL, USA, May 13-17.

Neti, C. V. and Senior, A. (1999). Audio-visual speaker recognition for video broadcast news. *In DARPA HUB4 Workshop*, pp. 139–142, Washington, DC, USA.

Neti, C., Potamianos, G., Luettin, J., Matthews, L., Glotin, H., Vergyri, D., Sison, J., Mashari, A. and Zhou, J. (2000). Audio-visual speech recognition. *Technical Report Workshop 2000, International Computer Science Institute, Center for Language and Speech Processing (CLSP)*, The Johns Hopkins University, Baltimore, MD, USA, October 12.

O'Shaughnessy, D. (1987). Speech Communications: Human and Machine, Series in Electrical Engineering ed. USA: Addison-Wesley Publishing Co.

Oudelha, M. and Ainon, R.N. (2010). HMM parameters estimation using hybrid Baum-Welch genetic algorithm. *International Symposium in Information Technology (ITSim)*, **2:** pp.542–545.

Pai, Y., Ruan, S., Shie, M., Liu, Y. (2006). A Simple and Accurate Color Face Detection Algorithm in Complex Background. *In ICME*, pp. 1545-1548.

Patterson, E.K., Gurbuz, S., Tufekci, Z. and Gowdy, J.N. (2002). Moving-talker speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus. *EURASIP Journal on Applied Signal Processing*, 11: pp.1189–1201.

Pentland, A. and Mase, K. (1989). Automatic lipreading by optical-flow analysis. Technical *Report VA189-8*, ITEJ.

Pérez, Ó, Piccardi, M. and García, J. (2007). Comparison between genetic algorithms and the Baum-Welch algorithm in learning HMMs for human activity classification**,** *Proceeding of EvoWorkshops'7*, pp.399–406.

Petajan, E. (1984). Automatic lipreading to enhance speech recognition, PhD. dissertation, Univ. Illinois at Urbana-Champagne.

Pigeon, S. and Vandendorpe. L. (1997). The M2VTS multimodal face database. *LNCS*, pp. 403–410.

Potamianos, G., Cosatto, E., Graf, H.P. and Roe, D.B. (1997). Speaker independent audio-visual database for bimodal ASR. In Benoît and Campbell (1997), pp. 65-68.

Potamianos, G., Verma, A., Neti, C. and Iyengar, G. (2000). A cascade image transform for speaker independent automatic speechreading. *In Proc. ICME*, pp. 1097-1100, New York, NY, USA.

Potamianos, G., Luettin, J. and Neti, C. (2001a). Hierarchical discriminant features for audio-visual LVCSR. *In Proc. ICASSP*, 1: pp. 165-168, Salt Lake City, UT, USA, May 7-11.

Potamianos, G., Neti, C., Iyengar, G. and Helmuth, E. (2001b). Large-vocabulary audio-visual speech recognition by machines and humans. *In Proc. 7th Eurospeech*, 2: pp. 1027-1030, Aalborg, Denmark, September 3-7.

Potamianos, G., Neti, C., Iyengar, G., Senior, A.W. and Verma, A. (2001c). A cascade visual front end for speaker independent automatic speechreading. Speech Technology, 4: pp. 193–208.

Rabiner, L. and Juang, B.H. (1993). Fundamentals of Speech Recognition. *Oxford University Press*.

Rao, R. and Mersereau, R. M. (1995). On merging hidden Markov models with deformable templates. *In Proc. ICIP*, 3: pp. 3556–3559, Washington, DC, USA.

Reisberg, D., McLean, J. and Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli », in Hearing by Eye : the psychology of lip-reading, B. Dodd et R. Campbell (eds.), Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp.97-114.

Revéret, L. (1999). Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole. *Thèse de doctorat*, de l'institut national polytechnique de Grenoble.

Robert-Ribes, J., Piquemal, M., Schwartz, J. L. and Escudier, P. (1996). Speechreading by Man and Machine: Models, Systems and Applications. chapitre Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition, pp. 193–210. Springer-Verlag, NATO ASI Series, Berlin, Germany.

Rodomagoulakis, I. (2008). Feature Extraction Optimization and Stream Weight Estimation in Audio-Visual Speech Recognition.*Phd thesis from Technical University of Crete*.

Rogozan, A., Deléglise, P. and Alissali, M. (1996). Intégration asynchrone des informations auditives et visuelles dans un système de reconnaissance de la parole », Actes des 21èmes Journées d'Etudes sur la Parole, Avignon, pp. 359-362.

Rogozan, A. (1999). Étude de la fusion des données hétérogènes pour la reconnaissance automatique de la parole audiovisuelle. *Thèse de doctorat*, Université d'Orsay - Paris XI.

Sánchez, U.R. (2000). Aspects of facial biometrics for verification of personal identity. Ph.D. thesis, University of Surrey, Guilford, UK.

Sanderson C. and Paliwal, K. (2002). Polynomial features for robust face authentication. *In proceedings of International Conference on Image Processing.*

Schwartz, J.-L., Robert-Ribès, J. and Escudier, P. (1998). Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. *chapitre Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception*, pp. 85–108. Psychology Press, Hove, UK.

Schwartz, J.-L. (2002). Traitement automatique du langage parlé 2: reconnaissance de la parole. *chapitre La parole multimodale: deux ou trois sens valent mieux qu'un*, pp. 141–178. Hermes, Paris.

Schwartz, J.-L. (2004). La parole multisensorielle: Plaidoyer, problèmes et perspectives. *In Actes des XXVème Journées d'Etude sur la Parole (JEP)*, pp. 11–17, Fès, Maroc.

Silsbee, P.L. and Su, Q. (1996). NATO ASI: Speechreading by Humans and Machines. *chapitre Audiovisual sensory integration using hidden Markov models*, pp. 489–495. Springer-Verlag.

Senior, A. W., (1999). Face and feature finding for a face recognition system. *In Proc. 2nd AVBPA*, pp. 154–159, Washington, DC, USA, March 22-23.

Shdaifat, I., Grigat, R. R. and Luetgert, S. (2001). Viseme recognition using multiple feature matching. *In Proc. 7th Eurospeech*, 4: pp. 2431–2434, Aalborg, Denmark, September 3-7.

Shing-Tai, P., Ching-Fa, C. and Jian-Hong Z. (2010). Speech Recognition via Hidden Markov Model and Neural Network Trained by Genetic Algorithm. *Ninth International Conference on Machine Learning and Cybernetics*. Qingdao, 11-14 July.

Sobottka, K., and Pitas, I. (1996). Segmentation and tracking of faces in color images, Automatic face and gesture recognition, pp. 236–241.

Stevens, S.S., Volkman, J. and Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. Proc. of JASA, 3(8): pp. 185–190.

Stork, D.G. (1997). HAL's Legacy. 2001's Computer as Dream and Reality. MIT Press, Cambridge, MA, USA.

Sumby, W.H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, 26, pp. 212-215.

Summerfield, Q. (1979). Use of visual information for phonetic perception, *Phonetica*, 36: pp. 314-331.

Summerfield, Q. (1983). Audio-visual speech perception, lipreading and artificial stimulation. Hearing Science and Hearing Disorders, pp. 131–182.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visuel speech perception, in Hearing by Eye: The psychology of lipreading, B. Dodd and R. Campbell, eds.

Summerfield, Q., MacLeod A., McGrath M. and Brooke M. (1989). Lips, teeth, and the benefits of lipreading, *in Handbook of Research on Face Processing*, A.W. Young and H.D. Ellis (eds.), Elsevier Science Publishers, pp. 223-233.

Taboada, J., Feijoo, S., Balsa, R. and Hernandez, C. (1994). Explicit estimation of speech boundaries. *IEEE Proc. Sci. Meas. Technol*., 141: pp. 153-159.

Teissier, P., Robert-Ribès, J. and Schwartz, J.-L. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6): pp. 629–642.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), IEEE Computer Society Press, Jauai, Hawaii, December 8-14.

Waibel, A. and Lee, K.-F. (1990). (eds), Readings in Speech Recognition, San Mateo, CA: Morgan Kaufmann.

Walden, B. E., Prosek, A. and Montgomery (1977). Effect of training on the visual recognition of consonants, Journal of Speech and Hearing Research, 20: pp. 130-145.

Wark, T. & Sridharan, S. (1998). An approach to statistical lip modelling for speaker identification via chromatic feature extraction, in International Conference on Pattern Recognition, pp. 123-125.

Whalen D.H. (1990). Coarticulation is largely planned, Journal of Phonetics, 18(1), pp. 3-35.

Wojdel J.C. and Rothkrantz. L.J.M. (2001a). Robust video processing for lipreading applications. *In Proc. 6th Euromedia*, pp. 195-199, Valencia, Spain, April 18-20.

Wojdel J.C. and Rothkrantz. L.J.M. (2001b). Using aerial and geometric features in automatic lip-reading. *In Proc. 7th Eurospeech*, 4: pp. 2463-2466, Aalborg, Denmark, September 3-7.

Wolpert, D.H., and Macready, W.G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), pp. 67-82.

Wright, A.H. (1991). Genetic algorithms for real parameter optimization. In *Proceeding of the Foundation Of Genetic Algorithms*. FOGA.

Xue-ying, Z., Yiping, W. and Zhefeng, Z. (2007). A Hybrid Speech Recognition Training Method for HMM Based on Genetic Algorithm and Baum Welch Algorithm. *IEEE 2nd International conference on Innovative Computing, Information and Control (ICICIC'07),* pp.572.

Yang, J. and Waibel, A., (1996). A real-time face tracker. *In: Proc. 3rd IEEE Workshop on Application of Computer Vision*. pp. 142-147.

Yang, C., Soong, F.K. and Lee, T. (2007). Static and dynamic spectral features: their noise robustness and optimal weights for ASR. IEEE Trans. on ASSP, 3(15): pp. 1087-1097.

Young, S., Evermann, G., Gale, M., Hain, s.T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2006). The HTK Book (for HTK version 3.4). *Cambridge University Engineering Department, Ed.*

Zemlin, W.R. (1968). Speech and Hearing Science: Anatomy and Physiology, New Jersey, Prentice-Hall.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *Proc. of JASA*, 2(33): pp. 248.

Notations

AAM	Active Appearance Model
ACP	Analyse en Composantes Principales
ASR	Automatic Speech Recognition
AVASR	Audio-Visual Automatic Speech Recognition
BW	Baum-Welch algorithm
DCT	Discrete Cosine Transform
DI	Direct Integration
DWT	Discrete Wavelet Transform
FAP	Facial Animation Parameters
FCC	Face Color Classifier
FLMP	Fuzzy-Logical Model of Perception
HMM	Hidden Markov ModelS
ID	Identification Directe
ICP	Institut de la Communication Parlée
IFCC	Individuel Face Color Classifier
IS	Identification Séparée
GA	Genetic Algorithm
GFCC	General Face Color Classifier
GMM	Gaussian Mixture Model
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coding
LUT	Look-Up Table
MFCC	Mel-scaled Frequency Cepstral Coefficients
MLLT	Maximum Likelihood Linear Transform
MMI	Maximum Mutual Information
MSA	Multiscale Spatiale Analysis
PLP	Perceptual Linear Predictive
RAP	reconnaissance automatique de la parole
RASTA-PLP	RelAtive SpecTral Analysis-Perceptual Linear Predictive
ROI	Region Of Interest
SI	Separate Integration
SNR	Signal-to-Noise Ratio

Publications réalisées au cours de la thèse

Publications et conférences internationales :

Makhlouf A., Lazli, L. and Bensaker, B. (2012). Structure Evolution of Hidden Markov Models for an Automatic Speechreading. *Accepted paper for 7th International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, Lugano, Switzerland.

Makhlouf A., Lazli, L. and Bensaker, B. (2013a). Automatic Speechreading Using Genetic Hybridization of Hidden Markov Models. *In Proceeding of the IEEE World Congress on Computer and Information Technology (WCCIT'13)*, June 22-24, 2013, Sousse, Tunisia.

Makhlouf A., Lazli, L. and Bensaker, B. (2013b). Hybrid Hidden Markov Models and genetic algorithm for Robust Automatic visual speech recognition. *Journal of Information Technology Review (JITR)*, 4(3): pp. 105-114.

Makhlouf A., Lazli, L. and Bensaker, B. (2016). Structure Evolution of Hidden Markov Models for Audiovisual Arabic Speech Recognition. *International Journal of Signal and Imaging Systems Engineering, IJSISE, 9(1)*, pp.55–66.

Co-encadrement:

Master de recherche Reconnaissance des Formes et Intelligence Artificielle (Janvier 2015- Juin 2015)

Boukhatem Chemssedine, « *extraction des paramètres vocaux à l'aide d'une nouvelle méthode d'analyse acoustique* », un master pourtant sur la mise en œuvre de la méthode J-RASTA pour faire une extraction des paramètres acoustiques.