Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)
http://dare.uva.nl/document/174206

File ID	174206
Filename	Bibliography

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type	Dissertation
Title	Audiovisual fusion for speaker diarization
Author	A. Noulas
Faculty	Faculty of Science
Year	2010
Pages	vii, 167
ISBN	90-75691-06-8

FULL BIBLIOGRAPHIC DETAILS:
http://dare.uva.nl/record/340830

Copyright

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other then for strictly personal, individual use.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)

Bibliography

[1] 2001 hub5 english evaluation. http://www.ldc.upenn.edu/Catalog/ CatalogEntry.jsp?catalogId=LDC2002S13, 2001.

[2] Darpa effective, affordable, reusable speech-to-text (ears):. http://www.darpa.mil/ipto/programs/ears, 2004.

[3] Asimo the honda humanoid robot. http://world.honda.com/ASIMO/, 2009.

[4] Javier Ajmera and Chuck Wooters. A robust speaker clustering algorithm. In In Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding, pages 411–416, 2003.

[5] Jitendra Ajmera. Robust Audio Segmentation. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2004.

[6] Jitendra Ajmera, Herve Bourlard, and Itshak Lapidot. Improved Unknown-Multiple Speaker clustering using HMM. Technical report, IDIAP, 2002.

[7] Xavier Anguera. Xbic: Real-time cross probabilities measure for speaker segmentation. Technical report, International Computer Science Institute, 2019.

[8] Xavier Anguera and Javier Hernando. Evolutive speaker segmentation using a repository system. In International Conference on Speech and Language Processing, 2004.

[9] Xavier Anguera, Chuck Wooters, and Javier Hernando. Automatic cluster complexity and quantity selection: Towards robust speaker diarization. In MLMI, volume 4299 of Lecture Notes in Computer Science, pages 248–256. Springer, 2006.

[10] Raimo Bakis, Scott Chen, Ponani Gopalakrishnan, Ramesh Gopinath, Lazaros Polymenakos, and Martin Franz. Transcription of broadcast news shows with the ibm large vocabulary speech recognition system. In Proceedings of the Speech Recognition Workshop, pages 67–72, 1997.

[11] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean luc Gauvain. Improving speaker diarization. In Rich Transcription Workshop, 2004.

[12] Zohar Barzelay and Yoav Y. Schechner. Harmony in motion. In Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2007.

[13] Matthew J. Beal, Hagai Attias, and Nebojsa Jojic. Audio-video sensor fusion with probabilistic graphical models. In in Proceedings European Confernece Computer Vision, 2002.

[14] Homayoon S.A. Beigi and Stephane H. Mayes. Speaker, channel and environment change detection. In World Congress on Automation, 1998.

[15] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davis, and

Mark B. Sandler. A tutorial on onset detection in musical signals. IEEE Transactions on speech and signal processing, 13(5):1035–1047, 2015.

[16] Mathieu Ben, Michael Betser, Frederic Bimbot, and Guillaume Gravier. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms. In International Conference on Spoken Language Processing, 2004.

[17] Samy Bengio. An asynchronous hidden markov model for audio-visual speech recognition. In in Advances in Neural Information Processing Systems, NIPS 15. MIT Press, 2003.

[18] Yoshua Bengio. Learning deep architectures for ai. Technical Report 1312, Dept. IRO, Universite de Montreal, 2007.

[19] P. Bertelson, J. Vroomen, G. Wiegeraad, and B. de Gelder. Exploring the relation between McGurk interference and ventriloquism. In International Conference on Spoken Language Processing, pages 559–562, 1994.

[20] Jonas Beskow. Rule-based visual speech synthesis. In In Proceedings of Eurospeech '95, pages 299–302, 1995.

[21] Jeff Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute Berkeley, 1997.

[22] Stan Birchfield. An implementation of the kanade-lucas-tomasi feature tracker. http://www.ces.clemson.edu/ stb/klt/, 2007.

[23] Jean-François Bonastre, Perrine Delacourt, Corinne Fredouille, Teva Merlin, and Christian Wellekens. A speaker tracking system based on speaker turn detection for nist evaluation. In International Conference on Acoustics, Speech and Signal Processing, pages 1177–1180, 2000.

[24] Gary Bradski and Adrian Kaehler. Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Inc., 2008.

[25] Rodrigo F. Cádiz. A fuzzy-logic mapper for audiovisual media. Comput. Music J., 30(1):67–82, 2006. ISSN 0148-9267.

[26] J. Carletta. Announcing the ami meeting corpus. The ELRA Newsletter 1(1):3-5, January-March 2006.

[27] Miguel A. Carreira-Perpignan and Geoffrey E. Hinton. On contrastive divergence learning. In Artificial Intelligence and Statistics, 2005.

[28] Mauro Cettolo and Michele Vescovi. Efficient audio segmentation algorithms based on the bic. In International Conference on Speech and Language Processing, pages 537–540, 2003.

[29] Neal Checka, Kevin Wilson, Michael Siracusa, and Trevor Darrell. Multiple person and speaker activity tracking with a particle filter. In International Conference on Acoustics, Speech and Signal Processing. IEEE, May 2004.

[30] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, and P. Olsen. Automatic transcription of broadcast news. Speech Communications, 37(1-2):69–87, 2002.

[31] Tsuhan Chen. Audiovisual speech processing. Signal Processing Magazine, 1(18): 19–21, January 2001.

[32] Yunqiang Chen and Yong Rui. Real-time speaker tracking using particle filter sensor fusion. In Proceedings of the IEEE, pages 485–494, 2004.

[33] Scott S. Chenn and Ponani Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. In IEEE Acoustics, Speech, and Signal Processing, volume 2, pages 645–648, 1998.

[34] Scott S. Chenn and Ponani Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[35] Garrison W. Cottrell, Peter Muncro, and David Zipser. Image compression by back propagation: An example of external programming. Ics rep 8702, Institute for Cognitive Science, University of California, 1987.

[36] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In In Workshop on Statistical Learning in Computer Vision, European Conference Computer Vision, pages 1–22, 2004.

[37] Sébastien Cuendet, Dilek Z. Hakkani-Tür, and Elizabeth Shriberg. Automatic labeling inconsistencies detection and correction for sentence unit segmentation in conversational speech. In MLMI, volume 4892 of Lecture Notes in Computer Science, pages 144–155, 2007.

[38] Ross Cutler and Larry Davis. Look who's talking: Speaker detection using video and audio correlation. In in IEEE International Conference on Multimedia and Expo, pages 1589–1592, 2000.

[39] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz, Ivan Tashev, Li wei He, Alex Colburn, Zhengyou Zhang Zicheng Liu, and Steve Silverberg. Distributed meetings: A meeting capture and broadcasting system. In ACM Multimedia, 2002.

[40] Chris Dance, Jutta Willamowski, Lixin Fan, Cedric Bray, and Gabriela Csurka. Visual categorization with bags of keypoints. In ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

[41] Trevor Darrell, John W. Fisher III, William T. Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In Advances in Neural Information Processing Systems 13, pages 772–778. MIT Press, 2000.

[42] Trevor Darrell, John W. Fisher III, and Paul Viola. Audiovisual segmentation and the cocktail party effect. In International Conference on Multimodal Interfaces, pages 32–40, 2000.

[43] Trevor Darrell, John W. Fisher III, and Paul Viola. Audiovisual segmentation and the cocktail party effect. In International Conference on Multimodal Interfaces, pages 32–40, 2000.

[44] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 28(4):357–366, 1980.

[45] Perrine Delacourt, David Kryze, and Christian J. Wellekens. Distbic: A speaker-based segmentation for audio data indexing. In Speech Communication, pages 111–

126, 2000.

[46] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

[47] Daniel P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2010. online web resource.

[48] Gwenn Englebienne, Tim Cootes, and Magnus Rattray. A probabilistic model for generating realistic lip movements from speech. In Advances in Neural Information Processing Systems 20, pages 401–408, Cambridge, MA, 2008. MIT Press.

[49] David C. Van Essen. Organization of visual areas in macaque and human cerebral cortex. he Visual Neurosciences, pages 507–521, 2004. MIT Press.

[50] Jonathan G. Fiscus, Jerome Ajot, and John S. Garofolo. The rich transcription 2007 meeting recognition evaluation. pages 373–389, 2008.

[51] Waka Fujisaki and Shin'ya Nishida. Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. Experimental Brain Research, 166 (3-4):455–464, October 2014.

[52] M. J. F. Gales and S. J. Young. The theory of segment hidden markov models. Technical report, University of Cambridge, June 1993.

[53] Rashmi Gangadharaiah, Balakrishnan Narayanaswamy, and Narayanaswamy. A novel method for two-speaker segmentation. In International Conference on Acoustics, Speech and Signal Processing, 2004.

[54] Daniel Gatica-Perez, Guilliame Lathoud, Jean-Marc Odobez, and Iain McCowan. Multimodal multispeaker probabilistic tracking in meetings. IDIAP-RR 66, IDIAP, 2004.

[55] Zoubin Ghahramani, Michael I. Jordan, and Padhraic Smyth. Factorial hidden markov models. In Machine Learning. MIT Press, 1997.

[56] Iyengar Giridharan, Nock Harriet J., and Neti Chalapathy. Audio-visual synchrony for detection of monologues in video archives. In International Conference on Multimedia and Expo, pages 329–332. IEEE Computer Society, 2003.

[57] Herbert Gish, Man-Hung Siu, and Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. In International Conference on Acoustics, Speech and Signal Processing, pages 873–876, 1991.

[58] Vishwa Gupta, Gilles Boulianne, Patrick Kenny, Pierre Ouellet, and Pierre Dumouchel. Speaker diarization of french broadcast news. In International Conference on Acoustics, Speech, and Signal Processing. IEEE Computer Society, 2008.

[59] John Hershey and Javier Movellan. Using audio-visual synchrony to locate sounds. In Advances in Neural Information Processing Systems 12, pages 813–819. MIT Press, 1999.

[60] Geoffrey E. Hinton. Training products of experts by minimizing contrastive fdivergence. Technical Report GCNU TR 2000-004, Gatsby Computational Neuroscience Unit University College London, 2000.

[61] Geoffrey E. Hinton, S. Onsidero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18(7):1527–1554, 2006.

[62] Geoffrey E. Hinton and Terrence J. Sejnowski. Optimal perceptual inference. In Conference on Computer Vision and Pattern Recognition, pages 448–453. IEEE Computer Society, 1983.

[63] Marijn Huijbregts, Chuck Wooters, and Roeland Ordelman. Filtering the unknown: Speech activity detection in heterogeneous video collections. In INTERSPEECH, pages 2925–2928, 2007.

[64] Jeih-Weih Hung, Hsin-Min Wang, and Lin-Shan Lee. Automatic metric-based speech segmentation for broadcast news via principal component analysis. In International Conference on Spoken Language Processing, pages 121–124, 2000.

[65] Nock Harriet J., Iyengar Giridharan, and Neti Chalapathy. Multimodal processing by finding common cause. Communications of the ACM, 47(1):51–56, 2004.

[66] John W. Fisher III and Trevor Darrell. Probabalistic models and informative subspaces for audiovisual correspondence. In European Conference on Computer Vision Part III, pages 592–603, London, UK, 2002. Springer-Verlag.

[67] S. E. Johnson. Who spoke when? - automatic segmentation and clustering for determining speaker turns. Eurospeech, 5:2211–2214, 1999.

[68] S. E. Johnson and Phill C. Woodland. Speaker clustering using direct maximisation of the mllr-adapted likelihood. In International Conference on Spoken Language Processing, pages 1775–1779, 1998.

[69] B.H. Juang and Lawrence R. Rabiner. A probabilistic distance measure for hidden markov models. AT&T Technical Journal, 1(64), 1985.

[70] Pearl Judea. Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference. Morgan Kaufmann, September 1988.

[71] Brigitte Zellner Keller. auses and the temporal structure of speech. In Fundamentals of speech synthesis and speech recognition, pages 41–62, 1994.

[72] Thomas Kemp, Michael Schmidt, Martin Westphal, and Alex Waibel. Strategies for automatic segmentation of audio data. In International Conference on Acoustics, Speech, and Signal Processing, pages 1423–1423. IEEE Computer Society, 2000.

[73] Einat Kidron, Yoav Y. Schechner, and Michael Elad. Pixels that sound. In Conference on Computer Vision and Pattern Recognition, pages 88–95. IEEE Computer Society, 2013.

[74] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In International Conference on Acoustics, Speech, and Signal Processing, pages 3089–3092. IEEE Computer Society, 1999.

[75] Francis Kubala, Hubert Jin, Yspyros Matsoukas, Long Nguyen, Rich Schwartz, and John Makhoul. The 1996 bbn byblos hub-4 transcription system. In In Proc. of DARPA Speech Recognition Workshop, pages 90–93, 1997.

[76] Solomon Kullback and Richard A. Leibler. On information and sufficiency. Annals of Mathematical Statistics, 22:79–86, 1951.

[77] Kornel Laskowski and Tanja Schultz. Modeling vocal interaction for segmentation in meeting recognition. In Andrei Popescu-Belis, Steve Renals, and Hervé Bourlard, editors, MLMI, volume 4892 of Lecture Notes in Computer Science, pages 259–270, 2007.

[78] David G. Lowe. Object recognition from local scale-invariant features. In International Conference on Computer Vision, pages 1150–1157. IEEE Computer Society, 1999.

[79] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60:91–110, 2004.

[80] Jean luc Gauvain, Lori Lamel, and Gilles Adda. Partitioning and transcription of broadcast news data. In ICSLP'98, pages 1335–1338, 1998.

[81] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with application to stereo vision. In Internation Conference on Artificial Intelligence, pages 121–130, 1981.

[82] Jürgen Lüttin, Neil A. Thacker, and Steve W. Beet. Speaker identification by lipreading. In International Conference of Spoken Language Processing, pages 62–65, 1996.

[83] Harry McGurk and John McDonald. Hearing lips and seeing voices. Nature, 1(264): 746–748, 1976.

[84] Sylvain Meignier, Jean françois Bonastre, and Stéphane Igounet. E-hmm approach for learning and adapting sound models for speaker indexing. In in Proc. Odyssey Speaker and Language Recognition Workshop, pages 175–180, 2001.

[85] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Laurent Besacier, and Jean-Francois Bonastre. Benefits of prior acoustic segmentation for automatic speaker segmentation. In International Conference on Acoustics, Speech and Signal Processing, pages 397–400. IEEE, 2004.

[86] Thomas Minka. Expectation-maximization as lower bound maximization, 1999.

[87] Thomas Minka. Estimating a dirichlet distribution, 2000, revised 2003, 2009.

[88] Thomas Minka. Discriminative models, not discriminative training. Technical Report TR-2005-144, Microsoft Research Cambridge, October 2011.

[89] Xavier Anguera Miro. Robust Speaker Diarization for Meetings. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2006.

[90] Gianluca Monaci, Pierre Vandergheynst, and Friederich T. Sommer. Learning bimodal structure in audio-visual data. IEEE Transactions on Neural Networks, submitted, 2008.

[91] Dasrren C. Moore. The idiap smart meeting room. IDIAP-COM 07, IDIAP, 2002.

[92] Daniel Moraru, Mathieu Ben, and Guillaume Gravier. Experiments on speaker tracking and segmentation in radio broadcast news. In International Conference on Spoken Language Processing, 2019.

[93] Michalis K. Titsias Moray Allan and Christopher K. I. Williams. Fast learning of sprites using invariant features. In British Machine Vision Conference, 2005.

[94] Kazumasa Mori and Seiichi Nakagawa. Speaker change detection and speaker clustering using vq distortion for broadcast news speech recognition. In Systems and Computers in Japan, pages 413–416, 2001.

[95] Kevin Patrick Murphy. Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, University of Berkeley, California, USA, 2002.

[96] Radford M. Neal. Connectionist learning of belief networks. Artificial Intelligence, 56:71–113, 1992.

[97] Masafumi Nishida and Tatsuya Kawahara. Speaker indexing and adaptation using speaker clustering based on statistical model selection. In International Conference on Acoustics, Speech and Signal Processing, pages 353–356. IEEE, 2004.

[98] Athanasios Noulas and Ben J. A. Krose. On-line multimodal speaker diarization. In International Conference on Multimodal Interfaces, pages 351–358, 2007.

[99] Athanasios K. Noulas, Bas Terwijn, Gwenn Englebienne, and Ben J. A. Krose. Speaker detection for conversational robots using synchrony between audio and video. In Icair Workshop, International Conference on Robotics and Automation, page accepted, 2010.

[100] Athanasios K. Noulas, Nikos Vlassis, and Ben J. A. Krose. Cross entropy for learning in multi-modal streams. In Joint Workshop on MultiModal Interaction and Related Machine Learning Algorithms, 2007.

[101] Maja Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. Philosophical Transactions of Royal Society B, to appear, 2010.

[102] Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In ISCA Workshop on Speaker Recognition: A Speaker Oddyssey, 2001.

[103] David C. Plaut, Steven J. Nowlan, and Geoffrey E. Hinton. Experiments on learning by back propagation. CMU-CS 86–126, Carnegie-Mellon University, 1986.

[104] Zochova Polina and Radova Vlasta. Modified distbic algorithm for speaker change detection. In Interspeech, pages 3073–3076, Bonn, 2011.

[105] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior. Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91:1306–1326, 2003.

[106] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of the IEEE, pages 257–286, 1989.

[107] Douglas A. Reynolds and Pedro A. Torres-Carrasquillo. Approaches and applications of speaker diarization. In International Conference on Acoustics, Speech, and Signal Processing, pages 953–956. IEEE Computer Society, 2010.

[108] David E. Rumelhart, Geoggrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 1(323):533–536, 1986.

[109] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In Artificial Intelligence and Statistics, page to appear, 2009.

[110] Ananth Sankar, Fuliang Weng, Andreas Stolcke, and Ramana Rao Gadde. The development of sri's 1997 broadcast news transcription system. In DARPA Broadcast News Transcription and Understanding Workshop, pages 91–96, 1998.

[111] Gideon Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6: 461–464, 1978.

[112] Xu Shao and Jon Barker. Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment. Speech Commun., 50(4):337–353, 2008.

[113] Jianhao Shi and Carlo Tomasi. Good features to track. In Conference on Computer Vision and Pattern Recognition, pages 593–600. IEEE Computer Society, 1994.

[114] Shi sian Cheng and Hsin min Wang. A sequential metric-based audio segmentation

method via the bayesian, 2003.

[115] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In In Proc. of DARPA Speech Recognition Workshop, pages 97–99, 1997.

[116] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodl. The cambridge university march 2005 speaker diarisation system. In INTERSPEECH, pages 2437–2440, 2016.

[117] Mang-Hung Siu, George Yu, and Herbert Gish. An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In International Conference on Acoustics, Speech and Signal Processing, pages 189–192. IEEE, 1992.

[118] Paul Smolensky. Information processing in dynamical systems: foundations of harmony theory. Parallel distributed processing: explorations in the microstructure of cognition, 1:194–281, 1986.

[119] Ilya Sutskever, Geoffrey E. Hinton, and Graham W. Taylor. The recurrent temporal restricted boltzmann machine. In NIPS, pages 1601–1608, 2008.

[120] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In Advances in Neural Information Processing Systems 19, pages 1345–1352. 2007.

[121] T.F.Cootes, G.J. Edwards, and C.J.Taylor. Active appearance models. IEEE transactions on Pattern Analysis and Machine Intelligence, 23(6):681–685, 2001.

[122] Bayes Thomas. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society, pages 330–418, 1763. Reprinted with biographical note by G.A. Barnard in Niometrika 45 1958.

[123] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University, April 1991.

[124] Alain Tritschler and Ramesh Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In Eurospeech, 1999.

[125] Andrea Vedaldi and Brian Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[126] Michele Vescovi, Mauro Cettolo, and Romeo Rizzi. A dp algorithm for speaker change detection. In Eurospeech 2003, pages 2997–3000, 2003.

[127] Paul Viola and Michael Jones. Robust real-time object detection. In International Journal of Computer Vision, 2001.

[128] Howard D. Wactlar, Alexander G. Hauptmann, and Michael J. Witbrock. News on-demand experiments in speech. In ARPA STL Workshop, 1996.

[129] Perrine Delacourt David Kryze Christian J. Wellekens. Detection of speaker changes in an audio document, 1999.

[130] Max Welling, Michal Rosen-Zvi, and Geoffrey Hinton. Exponential family harmoniums with an application to information retrieval. In Advances in Neural Information Processing Systems 17, pages 1481–1488. 2012.

[131] Chuck Wooters, James Fung, Barbara Peskin, and Xavier Anguera. Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. In In Rich Transcription 2004 Fall Workshop, 2004.

[132] Chuck Wooters and Marijn Huijbregts. The ICSI RT07s speaker diarization system. Lecture Notes in Computer Science, pages 509–519, 2008.

[133] Masahide Yamaguchi, Masaru Yamashita, and Shoichi Matsunaga. Spectral cross-correlation features for audio indexing of broadcast news and meetings. In INTERSPEECH, pages 613–616, 2005.

[134] S.J. Young and Sj Young. The htk hidden markov model toolkit: Design and philosophy. Entropic Cambridge Research Laboratory, Ltd, 2:2–44, 1994.

[135] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1):39–58, 2009.

[136] Xuan Zhu, Claude Barras, Lori Lamel, and Jean-Luc Gauvain. Multi-stage speaker diarization for conference and lecture meetings. Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, pages 533–542, 2008.