INTERSPEECH 2018 Turorial: Multimodal Speech and Audio Processing in Audio-Visual Human-Robot Interaction

List of References

Tutorial Slides: http://cvsp.cs.ntua.gr/interspeech2018

Petros Maragos and Athanasia Zlatintsi

Sunday, September 2, 2018, 14:00 - 17:30

1 Audio-Visual Perception and Fusion

[1] P. Aleksic and A. Katsaggelos. Audio-visual biometrics. *Proceedings of the IEEE*, 11:2025–2044, 2006.

[2] S. Escalera, J. Gonzalez, X. Baro, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, , and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proc. 15th ACM Int'l Conf. on Multimodal Interaction*, 2013.

[3] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR-16)*, pages 1933–1941, 2016.

[4] P.P. Filntisis, A. Katsamanis, and P. Maragos. Photo-realistic adaptation and interpolation of facial expressions using hmms and aams for audio-visual speech synthesis. In *Proc. Int'l Conf. on Image Processing (ICIP-2017)*, Beijing, China, Sep. 2017.

[5] P.P. Filntisis, A. Katsamanis, P. Tsiakoulis, and P. Maragos. Video-realistic expressive audio-visual speech synthesis for the greek language. *Speech Communication*, 95:137–152, Dec. 2017.

[6] A. Katsaggelos, S. Bahaadini, and R. Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9):1635–1653, 2015.

[7] A. Katsamanis, G. Papandreou, and P. Maragos. Face active appearance modeling and speech acoustic information to recover articulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):411–422, 2009.

[8] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[9] P. Maragos, P. Gros, A. Katsamanis, and G. Papandreou. Cross-modal integration for performance improving in multimedia: A review. In *in Multimodal Processing and Interaction: Audio, Video, Text, edited by P. Maragos, A. Potamianos and P. Gros, Springer-Verlag*, 2008.

[10] P. Maragos, A. Potamianos, and P. Gros. *Multimodal Processing and Interaction: Audio, Video, Text*. Springer-Verlag, New York, 2008.

[11] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

[12] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition via multiple hypotheses rescoring. *The Journal of Machine Learning Research*, 16(1):255–284, 2015.

[13] G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Thermos. Audio and visual modality combination in speech processing applications. In *S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Kruger, eds., The Handbook of Multimodal-Multisensor Interfaces, Vol. 1: Foundations, User Modeling, and Multimodal Combinations*. Morgan Claypool Publ., San Rafael, CA, 2017.

[14] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[15] A. Tsiami, A. Katsamanis, P. Maragos, and A. Vatakis. Towards a behaviorally-validated computational audiovisual saliency model. In *Proc. 41st IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-16)*, Shanghai, China, Mar. 2016.

[16] E. Tsilionis and A. Vatakis. Multisensory binding: is the contribution of synchrony and semantic congruency obligatory? *Current Opinion in Behavioral Sciences*, 8:7–13, 2016.

[17] A. Vatakis, P. Maragos, I. Rodomagoulakis, and C. Spence. Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *Journal Speech Lang Hear Res*, 2012.

[18] A. Vatakis and C. Spence. Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, 1111:134–142, 2006.

[19] A. Vatakis and C. Spence. Crossmodal binding: Evaluating the ?unity assumption? using audiovisual speech stimuli. *Attention, Perception, & Psychophysics*, 69(5):744–756, 2007.

[20] J. Wu, J. Cheng, et al. Bayesian co-boosting for multi-modal gesture recognition. *Journal of Machine Learning Research*, 15(1):3013–3036, 2014.

2 Audio-Visual HRI: Methodology and Applications in Assistive Robotics

[1] J. Broekens, M. Heerink, , and H. Rosendal. Assistive social robots in elderly care: A review. *Gerontechnology*, 8(2):203–275, 2009.

[2] G. Chalvatzaki, X.S. Papageorgiou, C.S. Tzafestas, and P. Maragos. Augmented human state estimation using interacting multiple model particle filters with probabilistic data association. In *Proc. IEEE Int'l Conf. on Robotics & Automation (ICRA-18)*, Brisbane, Australia, 2018.

[3] G. Chalvatzaki, G. Pavlakos, K. Maninis, X.S. Papageorgiou, V. Pitsikalis, C.S. Tzafestas, and P. Maragos. Towards an intelligent robotic walker for assisted living using multimodal sensorial data. In *Proc. Int'l Conf. on Wireless Mobile Communication and Healthcare (Mobihealth-14)*, pages 156–159. IEEE, 2014.

[4] A. Dometios, A. Tsiami, A. Arvanitakis, P. Giannoulis, X. Papageorgiou, C. Tzafestas, and P. Maragos. Integrated speech-based perception system for user adaptive robot motion planning in assistive bath scenarios. In *Proc. of the 25th European Signal Proc. Conf. - Workshop: "MultiLearn 2017 - Multimodal processing, modeling and learning for human-computer/robot interaction applications"*, Kos, Greece, Aug.-Sep. 2017.

[5] A.C. Dometios, X.S. Papageorgiou, A. Arvanitakis, C.S. Tzafestas, and P. Maragos. Real-time end-effector motion behavior planning approach using on-line point-cloud data towards a user adaptive assistive bath robot. In *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2017)*, pages 5031–5036. IEEE, 2017.

[6] E. Efthimiou, S.-E. Fotinea, T. Goulas, A.-L. Dimou, M. Koutsombogera, V. Pitsikalis, P. Maragos, and C. Tzafestas. The mobot platform–showcasing multimodality in human-assistive robot interaction. In *Proc. Int'l Conf. on Universal Access in Human-Computer Interaction*, pages 382–391. Springer, 2016.

[7] M. A. Goodrich and A. C. Schultz. Human-robot interaction: A survey. *Found. trends human-computer Interact.*, 1(3):203–275, 2007.

[8] A. Guler, N. Kardaris, S. Chandra, V. Pitsikalis, C. Werner, K. Hauer, C. Tzafestas, P. Maragos, and I. Kokkinos. Human joint angle estimation and gesture recognition for assistive robotic vision. In *Proc. European Conference on Computer Vision*, pages 415–431. Springer, 2016.

[9] R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu. Socially assistive robots in elderly care: A mixed-method systematic literature review. *Intl Jour. Human-Computer Interaction*, 30(5):369–393, 2014.

[10] N. Kardaris, V. Pitsikalis, E. Mavroudi, and P. Maragos. Introducing temporal order of dominant visual word sub-sequences for human action recognition. In *Proc. Int'l Conf. on Image Processing (ICIP-2016)*, pages 3061–3065. IEEE, 2016.

[11] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos. A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands. In *Proc. of the 2016 ACM on Multimedia Conf.*, pages 1169–1173. ACM, 2016.

[12] A. Katsamanis, V. Pitsikalis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition. In *The Handbook of Multimodal-Multisensor Interfaces*, pages 449–487. Association for Computing Machinery and Morgan & Claypool, 2017.

[13] A. Kotteritzsch and B. Weyers. Assistive technologies for older adults in urban areas: A literature review. *Cognitive Computation*, 8:299–317, 2016.

[14] P. Maragos, V. Pitsikalis, A. Katsamanis, N. Kardaris, E. Mavroudi, I. Rodomagoulakis, and A. Tsiami 2015. Multimodal sensory processing for human action recognition in mobility assistive robotics. In *Proc. IROS-2015 Workshop on Cognitive Mobility Assistance Robots*, Hamburg, Germany, Sep. 2015.

[15] E. Mordoch, A. Osterreicher, L. Guse, K. Roger, and G. Thompson. Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas*, 74:14–20, 2013.

[16] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition via multiple hypotheses rescoring. *The Journal of Machine Learning Research*, 16(1):255–284, 2015.

[17] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, A. Arvanitakis, and P. Maragos. A multimedia gesture dataset for human robot communication: Acquisition, tools and recognition results. In *Proc. Int'l Conf. on Image Processing (ICIP-2016)*, pages 3066–3070. IEEE, 2016.

[18] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos. Multimodal human action recognition in assistive human-robot interaction. In *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-16)*, pages 2702–2706. IEEE, 2016.

[19] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos. Room-localized spoken command recognition in multi-room, multi-microphone environments. *Computer Speech & Language*, 46:419–443, 2017.

[20] F. Rudzicz, R. Wang, M. Begum, , and A. Mihailidis. Speech interaction with personal assistive robots supporting aging at home for individuals with alzheimers disease. *ACM Trans. Access. Comput.*, 7(2):1–222, 2015.

[21] A. Zlatintsi, I. Rodomagoulakis, P. Koutras, A. C. Dometios, V. Pitsikalis, C. S. Tzafestas, and P. Maragos. Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot. In *Proc. 43rd IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-18)*, Calgary, Canada, Apr. 2018.

[22] A. Zlatintsi, I. Rodomagoulakis, V. Pitsikalis, P. Koutras, N. Kardaris, X. Papageorgiou, C. Tzafestas, and P. Maragos. Social human-robot interaction for the elderly: two real-life

use cases. In *Proc. of the 2017 ACM/IEEE Int'l Conf. on Human-Robot Interaction*, pages 335–336. ACM, 2017.

3 A-V Child-Robot Interaction

[1] T. Belpaeme, P. Baxter, R. Read, R. Wood, and et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[2] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos. Multi-view fusion for action recognition in child-robot interaction. In *Proc. Int'l Conf. on Image Processing (ICIP-18)*, Athens, Greece, Oct. 2018.

[3] P. Giannoulis, G. Potamianos, and P. Maragos. On the joint use of nmf and classification for overlapping acoustic event detection. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 2, page 90, 2018.

[4] J. Hadfield, P. Koutras, N. Efthymiou, G. Potamianos, C.S. Tzafestas, and P. Maragos. Object assembly guidance in child-robot interaction using rgb-d based 3d tracking. In *Proc. of 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS-2018)*, Madrid, Spain, Oct. 2018.

[5] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proc. on Human Robot Interaction (HRI-17)*, 2017.

[6] A. Potamianos, C. Tzafestas, E. Iosif, F. Kirstein, P. Maragos, K. Dauthenhahn, J. Gustafson, J.E. Ostergaard, S. Kopp, P. Wik, et al. Babyrobot—next generation social robots: Enhancing communication and collaboration development of td and asd children by developing and commercially exploiting the next generation of human-robot interaction technologies. In *Proc. of the Workshop on Evaluating Child-Robot Interaction (CRI) at the ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI)*, volume 495, 2016.

[7] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120, 2005.

[8] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, and G. Potamianos andP. Maragos. Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults. In *Proc. 43rd IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-18)*, Calgary, Canada, Apr. 2018.

[9] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos. Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA-18)*, Brisbane, Australia, May 2018.

4 Multimodal Saliency and Summarization

[1] C. Alain and L.J. Bernstein. From sounds to meaning: the role of attention during auditory scene analysis. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 16(5):485–489, 2008.

[2] D. Dimitriadis, P. Maragos, and A. Potamianos. On the effects of filterbank design and energy computation on robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1504–1516, 2011.

[3] M. Elhilali, J. Xiang, S.A. Shamma, and J.Z. Simon. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS biology*, 7(6):e1000129, 2009.

[4] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[5] J.B. Fritz, M. Elhilali, S.V. David, and S.A. Shamma. Auditory attention?focusing the searchlight on sound. *Current opinion in neurobiology*, 17(4):437–455, 2007.

[6] E.R. Hafter, A. Sarampalis, and L. Psyche. Auditory attention and filters. In *Auditory perception of sound sources*, pages 115–142. Springer, 2008.

[7] S. Haykin and Z. Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[8] J.F. Kaiser. On a simple algorithm to calculate the energy of a signal. In *Proc. Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP-90)*, pages 381–384. IEEE, 1990.

[9] E. M. Kaya and M. Elhilali. A temporal saliency map for modeling auditory attention. In *Proc. Conf. on Information Sciences and Systems (CISS-12)*, pages 1–6. IEEE, 2012.

[10] E. M. Kaya and M. Elhilali. Modelling auditory attention. *Phil. Trans. R. Soc. B*, 372(1714):20160101, 2017.

[11] E.M. Kaya and M. Elhilali. Investigating bottom-up auditory attention. *Frontiers in human neuroscience*, 8:327, 2014.

[12] C. Kayser, C.I. Petkov, M. Lippert, and N.K. Logothetis. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

[13] P. Koutras and P. Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015.

[14] P. Koutras, G. Panagiotaropoulou, A. Tsiami, and P. Maragos. Audio-visual temporal saliency modeling validated by fmri data. In *Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshops*, pages 2000–2010, 2018.

[15] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *Proc. Int'l Conf. on Image Processing (ICIP-15)*, pages 4361–4365. IEEE, 2015.

[16] P. Koutras, A. Zlatintsi, and P. Maragos. Exploring cnn-based architectures for multimodal salient event detection in videos. In *Proc. 13th IEEE Image, Video, and Multidimensional Signal Processing (IVMSP-18) Workshop*, Zagori, Greece, June 2018.

[17] P. Maragos, J.F. Kaiser, and T.F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing*, 41(10):3024–3051, 1993.

[18] G. Panagiotaropoulou, P. Koutras, A. Katsamanis, P. Maragos, A. Zlatintsi, A. Protopapas, E. Karavasilis, and N Smyrnis. Fmri-based perceptual validation of a computational model for visual and auditory saliency in videos. In *Proc. IEEE Int'l Conf on Image Processing (ICIP-16)*, pages 699–703. IEEE, 2016.

[19] T. Tsuchida and G.W. Cottrell. Auditory saliency using natural statistics. In *CogSci*, 2012.

[20] J. Wang, K. Zhang, K. Madani, and C. Sabourin. Salient environmental sound detection framework for machine awareness. *Neurocomputing*, 152:444–454, 2015.

[21] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32, 2008.

[22] A. Zlatintsi, E. Iosif, P. Maragos, and A. Potamianos. Audio salient event detection and summarization using audio and text modalities. In *Proc. European Signal Processing Conf. (EUSIPCO-15)*, pages 2311–2315. IEEE, 2015.

[23] A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos, and K. Pastra. Quality evaluation of computational models for movie summarization. In *Proc. Int'l Workshop on Quality of Multimedia Experience (QoMEX-15)*, pages 1–6. IEEE, 2015.

[24] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos. Cognimuse: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):54, 2017.

[25] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos. A saliency-based approach to audio event detection and summarization. In *Proc. European Signal Processing Conf. (EUSIPCO-12)*, pages 1294–1298. IEEE, 2012.

[26] E Zwicker and H Fastl. *Psychoacoustics Facts and Models Springer Heiderberg*. Springer, 2nd edition, 1999.

5 Audio-Gestural Music Synthesis

[1] C. Garoufis, A. Zlatintsi, and P. Maragos. A collaborative system for composing music via motion using a kinect sensor and skeletal data. In *Proc. Sound and Music Computing Conference (SMC-2018)*, Limassol, Cyprus, July 2018.

[2] M. Gleicher and N. Ferrier. Evaluating video-based motion capture. In *Proc. Computer Animation Conf. (CA-02)*, Switzerland, 2002.

[3] R. I. Godoy and M. Leman. *Musical Gestures: Sound, Movement, and Meaning*. New York: Routledge, 2010.

[4] A. Mulder. Virtual musical instruments: Accessing the sound synthesis universe as a performer. In *Proc. Brazilian Symposium on Computer Music*, 1994.

[5] T. Winkler. Making motion musical: Gesture mapping strategies for interactive computer music. In *Proc. Computer Music Conf.*, Bannf, Canada, 1995.

[6] A. Zlatintsi, P.P. Filntisis, C. Garoufis, A. Tsiami, K. Kritsis, M.A. Kaliakatsos-Papakostas, A. Gkiokas, V. Katsouros, and P. Maragos. A web-based real-time kinect application for gestural interaction with virtual musical instruments. In *Proc. Audio Mostly Conference (AM'18)*, Wrexham, United Kingdom, Sep. 2018.