ETSI TS 103 296 V1.1.1 (2016-08)

**Speech and Multimedia Transmission Quality (STQ);
Requirements for Emotion Detectors
used for Telecommunication Measurement Applications;
Detectors for written text and spoken speech**

Reference
DTS/STQ-236

Keywords
internet, quality

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Specification (TS) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

It is important to mention that there is a difference between the concepts of sentiment and emotion. In the first case, it is expressed after deep thought and uses a well-organized lexicon. In the second case, it highly depends on specific situations and is expressed by physiological responses. Emotions are considered as a strong feeling while sentiment is mental attitude caused by feeling [i.146].

Hereinafter, these two terms will be represented as emotions and detectors dealing with these terms - emotion detectors.

1 Scope

The present document specifies Classification of and Requirements for Emotion Detectors for telecommunications and the assessment of their performance and uncertainties.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

[1] ISO/IEC Guide 98-3:2008: "Uncertainty of measurement -- Part 3: Guide to the expression of uncertainty in measurement (GUM:1995).

[2] Recommendation ITU-T P.1401 (07/2012): "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models".

[3] Spiegel, M. (1998): "Theory and problems of statistics", McGraw Hill.

[4] Recommendation ITU-T P.800 (08/1996): "Methods for subjective determination of transmission quality".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] P. Ekman: "Universals and cultural differences in facial expressions of emotion" Nebraska symposium on motivation, pp. 207-282, 1972.

[i.2] P. Ekman, W. V Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer and M. Tomita: "Universals and cultural differences in the judgments of facial expressions of emotion", Journal of personality and social psychology, vol. 53, no. 4. pp. 712-7, 1987.

[i.3] R. W. Picard: "Affective Computing", Pattern Recognit., vol. 20, no. 321, p. 304, 1995.

[i.4] H. Atassi, M. T. Riviello, Z. Smékal, A. Hussain and A. Esposito: "Emotional Vocal Expressions Recognition Using the COST 2102 Italian Database of Emotional Speech" in Development of Multimodal Interfaces Active Listening and Synchrony, 2010, pp. 255-267.

[i.5] H. Binali, C. Wu and V. Potdar: "Computational Approaches for Emotion Detection in Text", 4th IEEE Int. Conf. Digit. Ecosyst. Technol. - Conf. Proc. IEEE-DEST 2010, DEST 2010, vol. 37, no. 5, pp. 498-527, 2010.

[i.6] S. Gupta, A. Mehra and Vinay: "Speech emotion recognition using SVM with thresholding fusion", 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015, pp. 570-574.

[i.7] Y. Sun, C. Quan, X. Kang, Z. Zhang and F. Ren: "Customer emotion detection by emotion expression analysis on adverbs", Inf. Technol. Manag., vol. 16, no. 4, pp. 303-311, 2015.

[i.8] Y. Baimbetov, I. Khalil, M. Steinbauer and G. Anderst-Kotsis: "Using Big Data for Emotionally Intelligent Mobile Services through Multi-Modal Emotion Recognition", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9102, A. Geissbühler, J. Demongeot, M. Mokhtari, B. Abdulrazak and H. Aloulou, Eds. Cham: Springer International Publishing, 2015, pp. 127-138.

[i.9] U. Krcadinac, P. Pasquier, J. Jovanovic and V. Devedzic: "Synesketch: An Open Source Library for Sentence-Based Emotion Recognition", IEEE Trans. Affect. Comput., vol. 4, no. 3, pp. 312-325, 2013.

[i.10] A. Neviarouskaya, H. Prendinger and M. Ishizuka: "Affect Analysis Model: novel rule-based approach to affect sensing from text" Nat. Lang. Eng., vol. 17, no. 01, pp. 95-135, Sep. 2011.

[i.11] S. Aman and S. Szpakowicz: "Identifying Expressions of Emotion in Text" in Text, Speech and Dialogue, vol. 4629, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 196-205.

[i.12] G. Valkanas and D. Gunopulos: "A UI Prototype for Emotion-Based Event Detection in the Live Web" Human-Computer Interact. Knowl. Discov. Complex, Unstructured, Big Data, vol. 7947 LNCS, pp. 89-100, 2013.

[i.13] S. Shaheen, W. El-Hajj, H. Hajj and S. Elbassuoni: "Emotion Recognition from Text Based on Automatically Generated Rules" in 2014 IEEE International Conference on Data Mining Workshop, 2014, pp. 383-392.

[i.14] C. Ma, H. Prendinger and M. Ishizuka: "Emotion Estimation and Reasoning Based on Affective Textual Interaction", Lecture Notes Comput. Sci. (including Subser. Lecture Notes Artif. Intell. and Lecture Notes Bioinformatics), vol. 3784 LNCS, pp. 622-628, 2005.

[i.15] D. T. Ho and T. H. Cao: "A High-Order Hidden Markov Model for Emotion Detection from Textual Data", Lecture Notes in Computer Science (Knowledge Management and Acquisition for Intelligent Systems), vol. 7457, 2012, pp. 94-105.

[i.16] F. Ren and C. Quan: "Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing" Inf. Technol. Manag., vol. 13, no. 4, pp. 321-332, 2012.

[i.17] W. H. Lin, T. Wilson, J. Wiebe and A. Hauptmann: "Which side are you on?: identifying perspectives at the document and sentence levels", Proc. Tenth Conf. Comput. Nat. Lang. Learn., pp. 109-116, 2006.

[i.18] S. K. D'Mello, S. D. Craig, J. Sullins and A. C. Graesser: "Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue", International Journal of Artificial Intelligence in Education, vol. 16, no. 1. pp. 3-28, 2006.

[i.19] S. D'Mello, N. Dowell and A. Graesser: "Cohesion relationships in tutorial dialogue as predictors of affective states", Front. Artif. Intell. Appl., vol. 200, pp. 9-16, 2009.

[i.20] E. Cambria, M. Grassi, A. Hussain and C. Havasi: "Sentic Computing for social media marketing" Multimed. Tools Appl., vol. 59, no. 2, pp. 557-577, 2012.

[i.21] P. Carvalho and M. J. Silva: "Clues for Detecting Irony in User-Generated Contents: Oh...!! It's 'so easy' ;-)", First Int'l CIKM Work. Top. Anal. Mass Opin., pp. 53-56, 2009.

[i.22] N. Guarino, D. Oberle and S. Staab: "Handbook on Ontologies", Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[i.23] S. Baccianella, A. Esuli and F. Sebastiani: "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", Proc. Seventh Int. Conf. Lang. Resour. Eval., vol. 0, pp. 2200-2204, 2010.

[i.24] C. Havasi, R. Speer and J. B. Alonso: "ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge", Proc. Recent Adv. Nat. Languges Process. 2007, pp. 1-7, 2007.

[i.25] A. Neviarouskaya, H. Prendinger and M. Ishizuka: "SentiFul: A Lexicon for Sentiment Analysis", IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 22-36, Jan. 2011.

[i.26] C. O. Alm, D. Roth and R. Sproat: "Emotions from text: Machine learning for text-based emotion prediction", Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05, 2005, no. October, pp. 579-586.

[i.27] Chunling Ma, A. Osherenko, H. Prendinger and M. Ishizuka: "A chat system based on emotion estimation from text and embodied conversational messengers (preliminary report)", Proceedings of the 2005 International Conference on Active Media Technology, 2005. (AMT 2005), no. i, pp. 546-548.

[i.28] C. Strapparava and A. Valitutti: "WordNet-Affect: an affective extension of WordNet", Proc. 4th Int. Conf. Lang. Resour. Eval., pp. 1083-1086, 2004.

[i.29] A. Esuli and F. Sebastiani: "SentiWordNet: A publicly available lexical resource for opinion mining", Proc. 5th Conf. Lang. Resour. Eval., vol. 6, pp. 417-422, 2006.

[i.30] A. Neviarouskaya, H. Prendinger and M. Ishizuka: "SentiFul: Generating a reliable lexicon for sentiment analysis", 3rd Int. Conf. Affect. Comput. Intell. Interact. Work., pp. 1-6, 2009.

[i.31] P. Ekman: "An argument for basic emotions", Cogn. Emot., vol. 6, no. 3, pp. 169-200, Jan. 1992.

[i.32] J. L. Tracy and D. Randles: "Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson and Panksepp and Watt", Emot. Rev., vol. 3, no. 4, pp. 397-405, 2011.

[i.33] C. E. Izard, D. Z. Libero, P. Putnam and O. M. Haynes: "Stability of emotion experiences and their relations to traits of personality", J. Pers. Soc. Psychol., vol. 64, no. 5, pp. 847-860, 1993.

[i.34] J. T. Hancock, C. Landrigan and C. Silver: "Expressing emotion in text-based communication", Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07, 2007, pp. 929-932.

[i.35] E. Cambria, T. Benson, C. Eckl and A. Hussain: "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality", Expert Syst. Appl., vol. 39, no. 12, pp. 10533-10543, 2012.

[i.36] J. A. Russell: "A circumplex model of affect", J. Pers. Soc. Psychol., vol. 39, no. 6, pp. 1161-1178, 1980.

[i.37] J. Posner, J. A. Russell and B. S. Peterson: "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development and psychopathology", Dev. Psychopathol., vol. 17, no. 03, pp. 715-734, Sep. 2005.

[i.38] N. A. Remington, L. R. Fabrigar and P. S. Visser: "Reexamining the circumplex model of affect", J. Pers. Soc. Psychol., vol. 79, no. 2, pp. 286-300, 2000.

[i.39] E. Cambria, A. Livingstone and A. Hussain: "The hourglass of emotions", Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7403 LNCS, pp. 144-157, 2012.

[i.40] A. Agrawal and A. An: "Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations", 2012 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., pp. 346-353, 2012.

[i.41] H. S. Ling, R. Bali and R. A. Salam: "Emotion detection using keywords spotting and semantic network IEEE ICOCI 2006", International Conference on Computing & Informatics 2006, pp. 1-5.

[i.42] B. Pang, L. Lee and S. Vaithyanathan: "Thumbs up?: sentiment classification using machine learning techniques", Proc. Conf. Empir. Methods Nat. Lang. Process., pp. 79-86, 2002.

[i.43] D. Wang, L. Shi, D. S. Yeung, P.-A. Heng, T.-T. Wong and E. C. C. Tsang: "Support vector clustering for brain activation detection", Med. Image Comput. Comput. Assist. Interv., vol. 8, no. Pt 1, pp. 572-579, 2005.

[i.44] R. Burget, J. Karásek and Z. Smékal: "Recognition of emotions in Czech newspaper headlines", Radioengineering, vol. 20, no. 1, pp. 39-47, 2011.

[i.45] W. Zheng and Q. Ye: "Sentiment classification of Chinese traveler reviews by support vector machine algorithm", 3rd Int. Symp. Intell. Inf. Technol. Appl. IITA 2009, vol. 3, pp. 335-338, 2009.

[i.46] A. McCallum and K. Nigam: "A Comparison of Event Models for Naive Bayes Text Classification", AAAI/ICML-98 Work. Learn. Text Categ., pp. 41-48, 1998.

[i.47] K.-M. Schneider: "On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification", Adv. Nat. Lang. Process., pp. 474-485, 2004.

[i.48] J. Huang, J. Lu and C. X. Ling: "Comparing naive Bayes, decision trees and SVM with AUC and accuracy", Third IEEE Int. Conf. Data Min., pp. 11-14, 2003.

[i.49] C. M. Bishop: "Pattern Recognition and Machine Learning", Springer-Verlag New York, ISBN 978-0-387-31073-2, 2006.

[i.50] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno and J. Caro: "Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning", IISA 2013 - 4th International Conference on Information, Intelligence, Systems and Applications, 2013, pp. 198-205.

[i.51] M. M. Itani, R. N. Zantout, L. Hamandi and I. Elkabani "Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes", 2nd Int. Conf. Adv. Comput. Tools Eng. Appl., pp. 192-197, 2012.

[i.52] S. Yoshida, J. Kitazono, S. Ozawa, T. Sugawara, T. Haga and S. Nakamura: "Sentiment analysis for various SNS media using Naïve Bayes classifier and its application to flaming detection", 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD), 2014, pp. 1-6.

[i.53] Z. Ghahramani: "An Introduction to Hidden Markov Models and Bayesian Networks", Int. J. Pattern Recognit. Artif. Intell., vol. 15, no. 01, pp. 9-42, Feb. 2001.

[i.54] B. Schuller, G. Rigoll and M. Lang: "Hidden Markov model-based speech emotion recognition", 2003 IEEE Int. Conf. Acoust. Speech, Signal Process. 2003. Proceedings. (ICASSP '03)., vol. 2, pp. 401-404, 2003.

[i.55] S. Rustamov, E. Mustafayev and M. a. Clements: "Sentiment analysis using Neuro-Fuzzy and Hidden Markov models of text", 2013 Proc. IEEE Southeastcon, pp. 1-6, 2013.

[i.56] E. C.-C. Kao, C.-C. Liu, T.-H. Yang, C.-T. Hsieh and V.W. Soo: "Towards Text-based Emotion Detection A Survey and Possible Improvements", 2009 International Conference on Information Management and Engineering, 2009, pp. 70-74.

[i.57] T. Mullen and N. Collier: "Sentiment Analysis using Support Vector Machines with Diverse Information Sources", Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. (EMNLP 2004), pp. 412-418, 2004.

[i.58] N. Chirawichitchai: "Sentiment classification by a hybrid method of greedy search and multinomial naïve bayes algorithm", 2013 Eleventh International Conference on ICT and Knowledge Engineering, 2013, pp. 1-4.

[i.59] X. Sun and C. Li: "Hybrid model based sentiment classification of Chinese micro-blog", 2014 International Conference on Audio, Language and Image Processing, 2014, pp. 358-361.

[i.60] D. S. Nair, J. P. Jayan, Rajeev R.R and E. Sherly: "Sentiment Analysis of Malayalam film review using machine learning techniques", 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015, pp. 2381-2384.

[i.61] P. C. R. Lane, D. Clarke and P. Hender: "On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data", Decis. Support Syst., vol. 53, no. 4, pp. 712-718, Nov. 2012.

[i.62] C. Quan and F. Ren: "A blog emotion corpus for emotional expression analysis in Chinese", Comput. Speech Lang., vol. 24, no. 4, pp. 726-749, Oct. 2010.

[i.63] France, D.J., Shiavi, R.G., Silverman, S. and Wilkes, M.: "Acoustical properties of speech as indicator of depression and suicidal risk", IEEE Trans. Biomed. Eng., 7, pp. 829-837.

[i.64] Mozziconacci, S.J.L. and Hermes, D.J.: "Expression of emotion and attitude through temporal speech variations", Proc. Int. Conf. On Spoken Language Processing (ICSLP 00), Beijing, vol. 2, pp. 373-378.

[i.65] Skinner, E. R.: "A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness", Speech Monographs, 1935.

[i.66] Friedhoff, A. J., Alpert, M., and Kurtzberg, R. L.: "An effect of emotion on voice", Nature, 193, 1962.

[i.67] Hecker M., Stevens, K., von Bismarck, G. and Williams, C. E.: "Manifestations of task-inducedstress in the acoustic speech signal", Journal of the Acoustical Society of America, 1968.

[i.68] Iida, A., Campbell, N., and Yasamura, M.: "Design and Evaluation of Synthesised Speech with Emotion", Journal of Information Processing Society of Japan, 40 (2), 1998.

[i.69] Douglas-Cowie, E, Cowie, R. and Campbell, N.: "Describing the emotional states that are expressed in speech", Special double issue of Speech Communication on "Speech and Emotion", 2003, vol 40 (1-2), pp 1- 257.

[i.70] Rahurkar, M. and Hansen, J. H. L: "Frequency band analysis for stress detection using a Teager energy operator based feature", Proc. Int. Conf. Spoken Language Processing (ICSLP '02). Vol. 3, 2002, pp. 2021-2024.

[i.71] Slaney, M. and McRoberts, G.: "A recognition system for affective vocalizations", Speech Communication 39, 367-384, Babyears 2003.

[i.72] Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O: "Combination of Machine Scores for Automatic Grading of Pronunciation Quality", Speech Comm., Vol. 30, pp. 121-130, 2000.

[i.73] Lee, C. M. and Narayanan, S. S.: "Toward detecting emotions in spoken dialogs", IEEE Trans. Speech and Audio Process. 13 (2), 293-303, 2005.

[i.74] K. Fischer: "Annotating emotional language data", Tech. Rep. 236, Univ. of Hamburg, 1999.

[i.75] Batliner, A., Hacker, C., Steidl, S., Noth, E., D' Archy, S., Russell, M. and Wong, M.: "You stupid tin box", children interacting with the AIBO robot: A crosslinguistic emotional speech corpus, Proc. Language Resources and Evaluation (LREC '04), Lisbon.

[i.76] Ken Bain: "Critical Thinking and Technology".

NOTE: Available at http://www.bestteachersinstitute.org/id98.html.

[i.77] Ververidis, D. and Kotropoulos, C.: "A review of emotional speech databases", Proc. Panhellenic Conference on Informatics (PCI), 2003, pp. 560-574.

[i.78] Plutchik, R.: "A psychoevolutionary theory of emotions. Social Science Information/sur les sciences sociales", Vol 21 (4-5), 1982, pp. 529-553. .

[i.79] Scherer, K.R.: "Expression of Emotion in Voice and Music", Journal of Voice, 1995 Lippincott-Raven Publishers, Philadelphia, vol. 9., No 3, pp. 235-248.

[i.80] Ververidis, D. and Kotropoulos, C.: "Automatic speech classification to five emotional states based on gender information", Proc. of 12th European Signal Processing Conference, Publisher IEEE, 2004, pp. 341-344.

[i.81] Ververidis, D. and Kotropoulos, C.: "Emotional speech recognition: Resources, features and methods", Speech communication, Publisher North-Holland, 2006, vol. 48, Nu.9, pp. 1162-1181.

[i.82] Petrusshin, V.A.: "Emotion in Speech: Recognition and Application to Call Centers", ResearchGate.

NOTE: Available at http://www.researchgate.net/publication/2611186.

[i.83] Polzin, T. and Waibel, A.: "Emotion-sensitive human-computer interfaces", Proc.of the ISCA workshop on Speech and Emotion, pp. 201-206, Newcastle, Northern Ireland, 2000.

[i.84] Syrový, V.: "Musical acoustic in Czech Hudební akustika", Ed. AMU, Prague, 2008. ISBN 978-80-7331-127-8.

[i.85] Tučková, J. and Šramka, M.: "ANN application in emotional speech analysis", Int.J. of Data Analysis Techniques and Strategies. 2012, vol. 4, no. 3/2012, p. 256-276. ISSN 1755-8050.

[i.86] Mahr, B.: "Die Informatik und die Logik der Modelle", Informatik Spektrum, 32(3): 228-249, 2009.

[i.87] Dupuis, K., Pichora-Fuller, M.K.: "Intelligibility of emotional speech in younger and older adults", Ear Hear, 2014, 35(6):695-707. Doi: 10,1097/AUD.0000000000000082.

[i.88] Pavlenko, A.: "Emotions and Multilingualism", Cambridge University Press. 2005, ISBN 0521843618.

[i.89] Orion Jones: "Different Languages Express Emotion Differently".

NOTE: Available at http://bigthink.com/ideafeed/different-languages-express-emotion-differently.

[i.90] Dewaele, J-M.: "Emotion in Nultiple languages", Publisher Palgrave Macmillan, Ebook, 2010, ISBN 978-1-4039-4316-3.

[i.91] Douglas-Cowie E., Cowie R. and Schroeder M.: "A New Emotion Database: Considerations, Sources and Scope", Proc. ISCA (ITWR) Workshop Speech and Emotion: A conceptual framework for research, pp. 39-44, Belfast, 2000.

[i.92] Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S. and Horton, D.: "Representation of prosodic and emotional features in a spoken language database", Proceedings of the 13th International Congress of Phonetic Sciences. Stockholm. 242-245, 1995.

[i.93] Campbell, N.: "Databases of Expressive Speech".

NOTE: Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.578&rep=rep1&type=pdf.

[i.94] Hansen., J.H.L.: SUSAS LDC99S78. Web Download. Philadelphia: Linguistic Data Consortium, 1999.

[i.95] Eckman, P.: "An argument for basic emotions", Cognition & Emotion 6, pp. 169-200, 1992.

[i.96] Ambrus D. C.: "Collecting and recording of an emotional speech Database", Technical Report, Faculty of Electrical Engineering and Computer Science, Institute of Electronics, University of Maribor.

[i.97] Ostermann J.: "Face animation in MPEG-4, in MPEG-4 Facial Animation" (I. S. Pandzic and R. Forchheimer, Eds.), pp.17-56, Chichester, U.K.: J. Wiley, 2002.

[i.98] Schroder, M.: "Emotional Speech Synthesis: A Review", Proceedings of Eurospeech, Aalborg, Denmark, 2001, s. 561-564.

[i.99] Cowie R., Douglas-Cowie E., Savvidou S., et al.: "Feeltrace: An instrument for recording perceived emotion in real time", Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, pp. 19-24, Belfast, 2000.

[i.100] McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M. and Stroeve, S.: "Approaching automatic recognition of emotion from voice: A rough benchmark", Proc. ISCA Workshop Speech Emotion, 2000, pp. 207-212.

[i.101] The Center for Spoken Language Research (CSLR), CU Kids' speech corpus.

[i.102] Linguistic Data Consurtium (LDC).

NOTE: Available at http://www.ldc.upenn.edu/.

[i.103] Pereira C.: "Dimensions of emotional meaning in speech", Proc. ISCA Workshop on Speech and Emotion: A conceptual framework for research, pp. 25-28, Belfast, 2000.

[i.104] Mehrabian, A. and Russell, J.A.: "An approach to environmental psychology", 1st ed., Cambridge, Mass.: MIT Press, 1974.

[i.105] Edgington M.: "Investigating the limitations of concatenative synthesis", Proc. Eurospeech 97, pp 593-596, Rhodes, Greece, September 1997.

[i.106] Polzin T. S. and Waibel A. H.: "Detecting emotions in speech", Proc. CMC, 1998.

[i.107] Petrushin, V. A.: "Emotion in speech recognition and application to call centers", Proc. ANNIE 1999, pp. 7-10.

[i.108] Fernandez R. and Picard R. W.: "Modeling drivers' speech under stress", Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, Belfast, 2002.

[i.109] Bavarian Archive for Speech Signals.

NOTE: Available at http://www.bas.uni-muenchen.de/Bas/.

[i.110] European Language Resources Association, (ELRA).

NOTE: Available at www.elra.info.

[i.111] Schiel F., Steininger Silke and Turk Ulrich: "The Smartkom Multimodal Corpus at BAS", Proc. Language Resources and Evaluation, Canary Islands, Spain, May 2002.

[i.112] Burkhardt F. and Sendlmeier, W.F.: "Verification of acoustical correlates of emotional speech using formant-synthesis", Proc. ISCA Workshop (ITRW) Speech and Emotion: A conceptual framework for research, Belfast, 2000.

[i.113] Kienast, M. and Sendlmeier, W. F.: "Acoustical analysis of spectral and temporal changes in emotional speech", ISCA Workshop on Speech and Emotion, Newcastle, UK, 2000, s. 92-97.

[i.114] Alter, K., Rank, E. and Kotz, S.A.: "Accentuation and emotions - Two different systems", Proc. ISCA Workshop on Speech and Emotion: A conceptual framework for research, Belfast, 2000.

[i.115] Wendt B. and Scheich H.: "The Magdeburger Prosodie-Korpus", Proc. Speech Prosody Conf., pp. 699-701, Aix-en-Provence, France, 2002.

[i.116] Schroder, M. and Grice, M.: "Expressing vocal effort in concatenative synthesis", Proc. 15th Int. Conf. Phonetic Sciences, Barcelona, Spain, 2003.

[i.117] Schroder, M.: "Experimental study of affect bursts", Proc. ISCA Workshop (ITRW) Speech and Emotion: A conceptual framework for research, pp. 132-137, Belfast, 2000.

[i.118] Scherer, K. R.: "Emotion effects on voice and speech: Paradigms and approaches to evaluation", Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, Belfast, 2000.

[i.119] Nakatsu, R., Solomides, A. and Tosa, N.: "Emotion recognition and its application to computer agents with spontaneous interactive capabilities", Proc. IEEE Int. Conf. Multimedia Computing and Systems, vol. 2, pp. 804-808, Florence, Italy, July 1999.

[i.120] Niimi Y., Kasamatu M.L., Nishimoto T. and Araki M.: "Synthesis of emotional speech using prosodically balanced VCV Segments", Proc. 4th ISCA tutorial and Workshop on research synthesis, Scotland, August 2001.

[i.121] Iida A., Campbell N., Iga S., Higuchi F. and Yasumura M.: "A speech synthesis system with emotion for assisting communication", Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, pp. 167-172, Belfast, 2002.

[i.122] Mozziconacci, S. J. L. and Hermes D. J.: "Expression of emotion and attitude through temporal speech variations", Proc. 2000 Int. Conf. Spoken Language Processing (ICSLP 2000), vol. 2, pp. 373-378, Beijing, China, 2000.

[i.123] Mozziconacci, S. J. L. and Hermes D. J.: "A study of intonation patterns in speech expressing emotion or attitude: production and perception", IPO Annual Progress Report 32, pp. 154-160, IPO, Eindhoven, The Netherlands, 1997.

[i.124] Montero, J. M., Gutierrez-Arriola, J., Colas, J., Enriquez, E. and Pardo, J. M.: "Analysis and modelling of emotional speech in Spanish", Proc. ICPhS'99, pp. 957-960, San Francisco 1999.

[i.125] Iriondo, I., Guaus, R. and Rodriguez, A.: "Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques", Proc. ISCA Workshop (ITRW) Speech and Emotion: A conceptual framework for research, pp. 161-166, Belfast, 2002.

[i.126] Engberg, I. S. and Hansen, A. V: "Documentation of the Danish Emotional Speech Database (DES)", Internal AAU report, Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark, September 1996.

[i.127] Amir, N., Ron, S. and Laor, N.: "Analysis of an emotional speech corpus in Hebrew based on objective criteria", Proc. ISCA Workshop (ITRW) Speech and Emotion: A conceptual framework for research, pp. 29-33, Belfast, 2000.

[i.128] Abelin, A. and Allwood, J.: "Cross linguistic interpretation of emotional prosody", Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, Belfast, 2000.

[i.129] Yu F., Chang E., Xu Y.Q. and Shum H.Y.: "Emotion detection from speech to enrich multimedia content", Proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp. 550-557, Beijing, China, October 2001.

[i.130] Makarova, V. and Petrushin, V. A.: "RUSLANA: A database of Russian Emotional Utterances", Proc. 2002 Int. Conf. Spoken Language Processing (ICSLP 2002), pp. 2041-2044, Colorado, USA, September 2002

[i.131] Constantini, G., Iadarola, I., Paoloni, A. and Todisco M.: "EMOVO Corpus: an Italian Emotional Speech Database", Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Publisher: European Language Resources Association (ELRA), ISBN 978-2-9517408-8-4.

[i.132] Clavel, Ch., Gael, R., Vasilescu, I., Devillers, L., Ehrette, T. and Sedogbo, C.: "The SAFE Corpus: illustrating extreme emotions in dynamic situations".

NOTE: Available at http://perso.telecom-paristech.fr/~grichard/Publications/LREC06_Clavel.pdf.

[i.133] Sigmund, M.: "Introducing the Database ExamStress for Speech under Stress", IEEE conf. Signal Processing Symposium, NORSIG 2006, Proceedings of the 7th Nordic.

[i.134] Hess, W.J.: "Pitch and voicing determination", Furui, S., Sondhi, M.M. (Eds.), Advances in Speech Signal Processing. Marcel Dekker, NY, 1992.

[i.135] Flanagan, J.L.: "Speech Analysis, Synthesis and Perception", 2nd Edition, Springer, NY, 1972.

[i.136] Nwe, T.L., Foo, S.W. and De Silva, L.C.: "Speech emotion recognition using hidden Markov models", Sppech Comm., 41, pp. 603-623.

[i.137] Rabiner, L.R. and Juang, B.H.: "Fundamentals of Speech Recognition", Prentice Hall, NY, 1993.

[i.138] Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S. and Horton, D.: "Representation of prosodic and emotional features in a spoken language database", Proceedings of the 13th International Congress of Phonetic Sciences. Stockholm. 1995, pp. 242-245.

[i.139] Petrushin, V.A.: "Emotion recognition in speech signal: experimental study development and application", Proc. of the 6th Int. Conference on Spoken Language Processing (ICSLP 2000) Beijing, China, 2000.

[i.140] Chul Min Lee, Shrikant and S. Narayanan: "Toward Detecting Emotions in Spoken Dialogs", IEEE Trans. On Speech and Audio Processing, vol.13, No.2, 2005.pp. 293-303, ISSN 1063-6676.

[i.141] Womack, B.D. and Hansen, J.H.L.: "Classification of speech under stress using target driven features", Speech Comm., vol. 20, pp. 131-150, 1996.

[i.142] Womack, B.D. and Hansen, J.H.L.: "N-channel hidden Markov models for combined stressed speech classification and recognition", IEEE Trans. Speech Audio Processing 7 (6), pp. 668-677, 1999.

[i.143] Fernandez, R. and Picard, R.: "Modeling drivers' speech under stress", Speech Comm. vol. 40, pp. 145-159, 2003.

[i.144] Roy, D. and Pentland, A.: "Automatic spoken affect analysis and classification", Proc. Int. Conf. Automatic Face Gesture Recognition. Killington, VT, 1996, pp. 363-367.

[i.145] Dellaert, F., Polzin, Th. and Waibel, A.: "Recognizing emotions in speech", ICSLP 96.

[i.146] Murray, I.R. and Arnott, J.L.: "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotions", J. Acoust. Society of America; 93 (2): 1097-1108.

[i.147] C. D. Broad: "Emotion and Sentiment", J. Aesthet. Art Crit., vol. 13, no. 2, pp. 203-214, 1954.

[i.148] Recommendation ITU-T P.800 (08/1996): "Methods for subjective determination of transmission quality".

[i.149] Gorayska B. and R. O. Lindsay: "The Roots of Relevance", Journal of Pragmatics 19, 301-323. Los Alamitos: IEEE Computer Society Press, 1993.

[i.150] Sheskin, David: "Handbook of Parametric and Nonparametric Statistical Procedures", CRC Press. ISBN 1584884401, 2004.

3 Definitions, symbols and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

curragator: muscle of the upper face which assists in expressing an emotion

diphone: unit of speech made up of two simple speech sounds known as phones

electroglottograph: device for measuring how much electricity flows across the larynx (it measures the variation in impedance to a very small electrical current between the electrode pair placed across the neck as the area of vocal fold contact changes during voicing

electromyogram: test that records the electrical activity of muscles

emotion detector: software tool capable of determining emotions present in a given input

laryngograph: instrument for recording the larynx movements in speech

MPEG-4: method of defining compression of audio and visual digital data

myogram: graphic representation of the phenomena (as velocity and intensity) of muscular contractions

PRAAT: free scientific computer software package for the analysis of speech in phonetics

SmartKom: multimodal dialogue system

synset: defined set of synonyms

3.2 Symbols

For the purposes of the present document, the following symbols apply:

BW_n	bandwidths of formants
E( )	Emotion vector
ESRM	percentage of Emotional Samples in Resulting Material
ESSM	percentage of Emotional Samples in Source Material
F	F-measure
F_0	fundamental frequency of the human voice
fn	percentage of false negative detected samples
F_n	formants (with n = 1, 2, 3, etc.)
fp	percentage of false positive detected samples
N	a number of stimuli considered in the comparison
R	Pearson correlation coefficient
T_0	pitch period $1/F_0$
tn	percentage of true negative detected samples
tp	percentage of true positive detected samples
wav	waveform audio file format
Xi	a value of the parameter characterizing for instance a strength of emotion carried by the text or speech sample
Yi	an estimated value of the parameter characterizing for instance a strength of emotion carried by the text or speech sample

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ACR	Absolute Category Rating
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASR	Automatic Speaker Recognition
ATR	Advanced Telecommunications Research
BAWE	British Academic Written English
BTW	By The Way
ATR Lab	Advanced Training & Research Laboratories
BT Labs	British Telecom research Laboratories
CPU	Central Processor Unit
CRF	Conditional Random Fields
DES	Danish Emotional Speech database
DSP	Digital Signal Processing
EEG	ElectroEncephaloGram
EISP	Emotion In Speech Project corpus
ELRA	European Language Resources Association
EMANN	Emotional database for ANN training
EmoDet	Emotion Detector
EMOVO	Italian Emotional Speech Database
ERR	Emotion Recognition Rule
ESRM	Emotional Samples in Resulting Material
ESSM	Emotional Samples in Source Material
FFM	Five Factor Model
FOAF	Friend Of A Friend
GUI	Graphical User Interface
HB	Hybrid approach Based

HMM	Hidden Markov Model
ISEAR	International Survey on Emotion Antecedents and Reactions
ITS	Intelligent Tutoring System
JST	Japan Science & Technology corporation
KB	Keyword Based
LDC	Linguistic Data Consortium
LOL	Laughing out loud
MFCC	Mel-Frequency Cepstral Coefficient
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLNN	Multi-Layer Neural Network
MPQA	Multi Perspective Question Answering
NBC	Naive Bayes Classifier
NEG	NEGative
NLP	Natural Language Processing
NSF/ITR	National Sanitation Foundation/Information Technology Research for national priorities
Micro-WNOp	Micro-WordNet-Opinion
MOS	Mean Opinion Score
OCC	Ortony, Clore & Collins emotion model
OMCS	Open Mind Common Sense
OMG	Oh My God
OMR	Ontology for Media Resources
PAD	Pleasure, Arousal, Dominance model
PANA	Positive Activation, Negative activation model
POS	POSitive
RAM	Random Access Memory
RB	Rule Based
RPS	Remaining Percentage of Samples
RUSSLANA	RUSSian LANguage Affective speech
SA	Sentiment Analysis
SAFE Corpus	Situation Analysis in a Fictional and Emotional Corpus
SES	Spanish Emotional Speech database
SIP	Social Information Processing
SOM	Self-Organizing Map

NOTE: A type of artificial neural network that is trained using unsupervised learning.

SUO-KIF	Standard Upper Ontology Knowledge Interchange Format
SUSAS	Speech Under Simulated and Actual Stress corpus
SVC	Support Vector Clustering
SVM	Support Vector Machines

NOTE: Supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

TESOL	Education program Teaching English to Speakers of Other Languages
TV	TeleVision
UK	United Kingdom
VAESS	Voice Attitudes and Emotions in Speech Synthesis
VAP	Valence, Arousal, Power model
VCV	Vowel Consonant Vowel

NOTE: One-syllable word has one consonant between two vowels.

WNA	WordNet-Affect®
WOZ	Wizard of OZ method

NOTE: Create the illusion of a working system by having a human operator.

www	world wide web

4 Emotion detectors for written text

4.1 Introduction

In recent years, recognition of emotions, sentiment analysis (SA), affect analysis and opinion mining have become a very popular topic for researches, the reason being that emotional state can influence the ways people behave. For example, the analysis of the emotional state is widely used by e-shops to analyse customers' reviews and identify their opinion about different products or services taking into account specific criteria. SA also can be used in social networks to analyse users' attitude in different situations (sadness, happiness and irony, among others).

Affect analysis is a method of analysing a given input and determining emotions of either text or speech data. There are various situations where sentiment analysis can be implemented for various purposes.

Since this is quite a popular subject, numerous approaches exist there and many algorithms which try to deal with this technology. However, at the time of the publication of the present document, none of them is completely developed and they have their inaccuracies and limitations.

Clause 4 of the present document collects information about current works, approaches and ideas related to this technology (including possible trials and demonstrations), investigates them and identifies features of emotion detectors. The information gathered from studied materials is presented in the following clauses, each addressing one key feature at a time.

4.2 Overview of emotion detectors

4.2.1 Introduction

Most of the emotion detectors and emotion recognition started with facial analysis [i.1] to [i.3], then continued to voice and developed to textual emotion recognition as well. Emotional detector is a powerful tool as it can be used for improving customer experience, removing inappropriate posts from social networks and discover opinions about products. Today there are software tools capable of recognizing an emotion and providing feedback in real-time. Examples of such technology can be the project from R. Picard [i.3], which allows real-time recognition of emotion from a person watching a video or browsing a web page. The purpose of this tool is to provide feedback to companies airing TV or commercials. It is possible to use it for different purposes, if there is a camera available.

Another example can be the emotion recognition from voice [i.4]. Such program can provide a convenient way how a call centre can evaluate its employees and can also help the operators to prevent drifting a conversation to an emotionally unpleasant state. It could offer automatically selection of problematic calls, where the caller appeared in a good mood at the beginning, but becomes upset at the end, which can finally help to improve the operator's performance. It can also help to select calls with a negative emotional potential to analyse them. The youngest area of this research is the textual emotion recognition which should be able to recognize emotions from text only. This tool has grown in importance in the last years, where computer mediated communication plays greater role than ever before. Textual recognition can be applied to statuses, posts from social networks, microblogs, blogs, online newspapers, forums and reviews. It can help with removal of inappropriate or abusive messages and detection of a negative content. Information gained from these sources can serve as a starting point to improving services, providing feedback, fixing errors and avoiding mistakes. Altogether these tools aim to improve user experience. Although mentioned programs target different areas of application, they share common core for emotion recognition.

4.2.2 Current approaches

Whether the goal is emotion recognition from image, voice or text, the means to achieve those are the same. Modern machine learning techniques made it possible to classify these emotions from images, but also from voice. Advanced signal processing algorithms are employed to do that. A summary of selected papers used in the present document is provided in table 1.

Similar machine learning approaches are taken in the area of textual emotion recognition. Even though there are simpler solutions taking advantage of key words or specific rules, machine learning is at the core of most methods. Binali, Gupta and Sun use techniques employing SVM classifier to determine the emotions in [i.5] to [i.7]. There are also more advanced and complex tools combining the rule based (keyword based) and machine learning methods. Baimbetov and Khalil use a hybrid method to determine emotions from instant messaging and speech to text convertor [i.8]. Their method is then compared to the results from voice emotion detector and Synesketch. Synesketch is a tool for graphical display of textual emotion information developed by Krcadinac [i.9] and is available online. His application uses a rule-based keyword spotting method to visualize human emotions in text. Rule-based linguistic approach is applied also in the work of Neviaroukaya and Prendinger in [i.10]. The authors created an Affect Analysis Model, which is capable of detecting emotions in blog posts and in instant messaging.

Instant messaging is a popular source of text for emotion recognition. Among other sources used are blog posts such as in the work of Aman [i.11]. Aman analysed blog posts on a sentence level. For emotion recognition a Naïve Bayes Classifier and Support Vector Machine was used. The task was to classify sentences into two categories, either emotional or not emotional.

Another source for textual emotion recognition is Twitter®. Valkanas used an algorithm with a decision tree classifier to recognize emotions from Twitter® in real time [i.12].

Twitter® was also used by Shaheen to train a classifier for recognition [i.13]. In his work, emotions were determined by an approach called Emotion Recognition Rule, which creates a complex rule-based structure for each input sentence. This structure is then presented to the nearest neighbour classifier which looks in the ERR annotated space. When the nearest neighbour is found, it delivers its emotion. The input sentence is also tested against WordNet and ConceptNet which are databases with important information related to words, their meaning and context.

Approaches mentioned earlier are summarized in table 1. As the table shows, there are three main ways to classify emotions. These methods can be distinguished as:

- Rule-based or keyword based.
- Machine learning based.
- Hybrid.

The emotion detector can be described as software, which has a textual input, algorithm and output related to used emotional model. Algorithm is connected to a word database and emotion model which allows it to determine emotion.

Table 1: Papers summary

Paper Name	Ontology	Lexicon/Corpora	Input granularity	Input textual data	Method of affect/sentiment analysis	Emotional model
[i.10]	WordNet	SentiFul®, WordNet-Affect®	fine-grained	Blogs, Instant Messaging	KB/RB	N/A
[i.14]	WordNet	OMCS, WordNet-Affect®	N/A	Chat system	N/A	Categorical
[i.5]	N/A	N/A	N/A	Intelligent tutor system	ML	Categorical
[i.11]	N/A	Blog posts annotated with Ekman's model, sentences	Sentences	Blog posts	ML	Categorical
[i.9]	WordNet	WordNet-based Lexicon, Lexicon of Emoticons, Abbreviations	Sentence-Level	N/A	KB/RB	Categorical
[i.6]	N/A	N/A	N/A	LDC Emotional Speech Database	ML	Categorical
[i.13]	N/A	Aman 2007, Twitter® - hashtag annotated	N/A	N/A	HB	Categorical
[i.12]	WordNet	WordNet-Affect®, Twitter® annotated by hashtags	N/A	Twitter®	ML	Categorical
[i.27]	WordNet	WordNet-Affect®	Word, Sentence	Chat program logs	KB/RB	Categorical
[i.15]	N/A	BAWE Corpus	N/A	ISEAR	KB/RB	Categorical
[i.16]	N/A	Ren-CECps	Sentence, fine-grained	Customer comments.	ML	Dimensional
[i.8]	N/A	N/A	Sentence-Level	Speech to Text, Instant messaging	HB	Categorical
[i.17]	N/A	Political articles	Sentence-Level, Document-Level	Political articles	HB	Categorical
[i.18]	N/A	AutoTutor Log Files	Sentence-Level, Chat exchange level	AutoTutor Log Files	ML	Categorical
[i.19]	N/A	AutoTutor Log Files, Video annotated by trained judges, Experiment participants cross annotation	Sentence-Level, Chat exchange level	AutoTutor Log Files	N/A	Categorical
[i.20]	FOAF, OMR, WNA, ConceptNet	LiveJournal® Blog posts, YouTube™ Videos (tagged like/dislike)	Blog Level, Video level	Blog posts	HB.	Dimensional

NOTE : Twitter®, LiveJournal® and YouTube™ are examples of a suitable products available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of these products.

4.2.3 Emotion detector description

The emotion detectors have many interpretations, yet most of them can be described as the input textual data, which is processed in an algorithm to produce emotional output. The top level view of such system is shown in figure 1. These systems have many different realizations for the algorithm, which may be using methods of machine learning or purely rule-based methods. One way or another, the algorithm shall operate with a certain emotional model. This model determines how the emotion will be displayed in the output and how it can be interpreted. Knowledge base, lexicon or ontology refers to a source of information used to train the classifier or to build a set of rules or keywords with appropriate emotion assigned to them. The emotion referenced in the knowledge base or lexicon shall also relate to the used emotional model properly. Each of these parts will be described in detail in clauses 4.3 to 4.8.

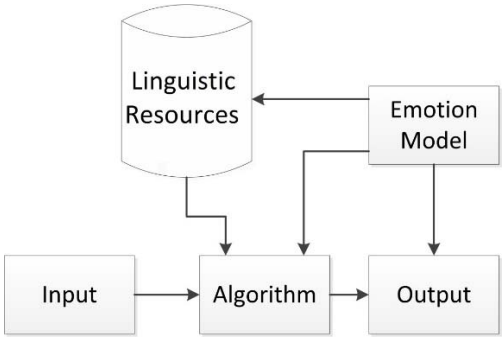


Figure 1: Emotion detector structure

4.3 Input

In this clause the main inputs are discussed, their sources, and requirements and recommendations for them. For the EmoDet model 2 types of input information were selected: text and meta-data, which can be derived from video or audio recordings.

For the first type of information (text), inputs can be arranged into 4 groups: words, sentences, paragraphs and articles.

Since the technology is still developing, grammatically incorrect sentences may cause problems with emotion recognition. There are approaches which can deal with this problem [i.10] using rule-based linguistic approach.

Abbreviations such as LOL, BTW, OMG and so on, and also emoticons should be addressed during pre-processing. For this case also exist approaches trying to find affect in above mentioned abbreviations but they are not mature [i.21].

It is also stated that emotions can be accompanied with vulgar lexicon and offensive terminology. This aspect should be dealt with as well [i.9].

As sources, following types of input data were acknowledged: typed by hand, or copied and pasted text, imported documents (.pdf, .doc, .xml and so on), e-mails and data from instant messaging services and posts from social networks and blogs.

4.4 Linguistic resources

4.4.1 Introduction

People use words to express happiness, frustration, despair, etc. but sometimes it is more complex to understand the underlying meaning of the words, for example when using irony or sarcasm. Another example is using idioms which hold a very specific meaning. This is a very demanding task. It becomes even harder when emotions are not presented explicitly using specific emotionally related words and are expressed using implicit cues (for example in education as noted by D'Mello [i.18] and [i.19]).

The diversity of words and their meaning is so complicated that, to interpret them, computers rely on many resources, which provide word's meaning, context, synonyms, use, origin, related words and other key points important to correct interpretation, information extraction and understanding. According to the type of information the resource holds, researchers talk about ontologies, corpora, lexicons, dictionaries, databases and knowledge bases. The terminology does not seem to be set exactly, some researchers use slightly different terms for similar bodies. For that reason clause 4.4.2 will cover the most frequent resources and their purpose.

4.4.2 Overview of resources

4.4.2.1 Description

Generally, authors use the term linguistic resource for any ontology, corpus, or lexicon.

Ontology is a way of describing things by their properties with a structured and specific manner. In different field ontology may have a slightly different meaning. From the top point of view ontology is a way of organizing objects to predefined structures. In terms of NLP, the ontologies are defined in terms of meaning, concepts and interrelations of words. More about this topic is described in [i.22]. Although it is a concept, some resources are regarded as ontologies.

The term corpus is a large body of natural language text, which is accompanied with a specific tag or information related to part of speech, parse tree and affective information. Corpus holds information based on its purpose. For that reason, it is often generated from a larger body and is extended with desired information. Corpora can be used for different tasks, for example training a classifier and validating one.

Lexicons describe a set of words by lexical categories to which they belong. The record also includes information about the meaning of the word and how it is used. It can be also accompanied by an affective meaning.

There are many linguistic resources belonging into one of the mentioned categories. The following resources are the most frequent.

4.4.2.2 WordNet

The most frequently used linguistic resource in the field of NLP for emotion recognition in text seems to be WordNet. Although it is described as a large lexical database of English, some consider it ontology as well. Nouns, verbs and other classes of words are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Conceptual, semantic and lexical relations link synsets together. There are 147 278 words at the time of writing tthe present document.

WordNet was explicitly mentioned in several papers [i.23] to [i.26], [i.10], [i.14], [i.9], [i.12] and [i.27] and was used for different purposes. For the purposes of textual emotion recognition, WordNet was used to create WordNet-Affect® [i.28], SentiWordNet [i.23], [i.29] and SentiFul® [i.25]. All three of them take a subset of synsets and extend the original records with appropriate information, e.g. affective labels, which suit emotion recognition. WordNet-Affect® was created mainly manually. SentiWordNet tried to exploit methods from machine learning to automatically assign emotional content to synsets. Similarly, SentiFul® is automatically generated lexicon with affective content.

There are other databases as well, for example SUMO, ConceptNet, Cyc™.

4.4.2.3 Suggested Upper Merged Ontology (SUMO)

The Suggested Upper Merged Ontology is free and owned by IEEE. Its domain ontologies form the largest formal public ontology in existence today. It has about 25 000 terms and about 80 000 axioms when all domain ontologies are combined. Main target areas are research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has been mapped to all of the WordNet. SUMO is written in the SUO-KIF language. Language generation templates are available for Hindi, Chinese, Italian, German, Czech and English.

4.4.2.4 Cyc™ database

It is a knowledge base of common sense. Its goal is to allow human-like reasoning for AI. Cyc™ is a formalized representation of fundamental knowledge of people. The representation is in a specific language called CycL. The knowledge base breaks down to terms, relations and assertions. There are specific APIs to allow programmers to use Cyc™. At the moment of publication of the present document, the APIs are for LISP and CycL. New assertions are continually expanding the knowledge base through a combination of automated and manual means.

NOTE : Cyc™ is an example of a suitable product available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of this product.

4.4.2.5 Open Mind Common Sense (OMCS)

In its nature is similar to Cyc™. It is also a knowledge base collecting fundamental human knowledge, which is taken as granted by humans. The knowledge was gained since 2000 from OMCS project. The project leveraged the power of social interactions on the Internet and people were filling in information in a questionnaire. This knowledge base is roughly ¼ of Cyc™ in size at the time of publication of the present document.

4.4.2.6 ConceptNet

It represents knowledge from OMCS as a directed graph. Nodes in the graph are concepts and labelled edges assertions.

4.4.2.7 Other databases

Other examples of linguistic resources used in NLP applications today are Ren-CECps (corpus in Chinese language) British Academic Written English (BAWE), International Survey on Emotion Antecedents and Reactions (ISEAR), Multi Perspective Question Answering (MPQA) and Micro-WordNet-Opinion (Micro-WNOp) Some linguistic resources, such as ISEAR, can be used directly for textual emotion recognition. Others resources need to be processed further to provide emotional meaning.

The above mentioned resources are then used in rule-based emotion recognition, or for training classifiers which will later be responsible for emotion recognition. Some sets are also used for the purpose of cross verification of already trained classifier or algorithm, which originally operated with different knowledge base. It is a common way of testing and expressing efficiency and precision of these algorithms.

To achieve trustworthy results, it is very important that the knowledge base itself does not contain many errors and that the records are related to proper emotional label. This is done with annotating the data.

4.4.3 Annotation

The data in a knowledge base is usually annotated by people to avoid any false annotations. Unfortunately this process is lengthy and expensive. Human annotators are slow and to annotate a large database of words it is necessary to employ many annotators. More importantly - to achieve a high level of precision, it is necessary to crosscheck annotators. At least certain number of annotators have to agree on the assigned label to allow it to enter the knowledge base. Quality and speed of achieved results depend on level of agreement. In case the annotators do not agree, the word have to be passed to another annotator or they have to discuss it to resolve the issue [i.25] and [i.30].

Table 2: Methods for annotating linguistic resources

Resource	Annotation
Twitter®, Weibo®, QQ®	hashtags
Facebook®	emoticons
WordNet-Affect®	manually
SentiWordNet®	automatically
SentiFul®	automatically
Instant messaging	emoticons
NOTE : Twitter®, Weibo®, QQ®, Facebook®, WordNet-Affect®, SentiWordNet®, SentiFul® and Instant messaging® are examples of a suitable products available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of these products.	

Therefore, there is a significant effort to simplify and speed up annotating process by employing automated ways of annotation. Some algorithms take advantage of already existing annotated knowledge bases. Others use internet resources. Means of annotating are summed up in table 2.

Thanks to modern means of communication, there is a big amount of data already available and practically annotated. Twitter® is one of the examples. It is a large microblogging network with more than 300 million tweets a day and a vast majority of tweets have hashtags. Hashtags are a very convenient way to be used for emotional annotation. Specific hashtags are related only to certain emotions and therefore the algorithm can focus on these hashtags only. With the body of the tweet is then possible to train a classifier or to create a new database with annotated sentences or words.

The same can be said about instant messaging or Facebook® statuses. Here the most important role is played by emoticons. The body of the message is annotated using an emoticon, which reflects a certain emotional state. The emoticons can be also used for Twitter®.

Despite the advantages of annotating algorithmically, using manual annotators is still the most reliable method.

4.5 Emotion models

4.5.1 Introduction

The effort to analyse human emotions started a long time ago and it is still an ongoing process. The first person who tried to deal with emotions was Cicero. He determined four basic emotion categories: "metus" (fear), "aegritudo" (pain), "libido" (lust), "laetitia" (pleasure). Discrete categories with labels reflecting emotions became the first way of assessing emotions. Charles Darwin proposed an evolutionary approach by stating that emotions are a product of evolution and there should be a limited number of basic emotions shared across cultures and animals. Ekman found evidence for a group of six basic emotions and since then there were other researchers taking this approach and creating several emotional categories according to their research. Nevertheless, Ekman's basic six emotions remain the most popular ones.

Emotions can be classified according to following criteria:

- Categorical.
- Dimensional.

Categorical means dividing emotions into several groups and assigning the emotion into a proper discrete group. These methods are based on works of Ekman, who introduced this concept [i.2].

Dimensional methods consider emotions as a place in two or even multi-dimensional space, where the emotion is expressed as a point in the plane or space defined by the variables taken into account.

To determine the emotion from a textual input, it is crucial to have some framework to use. In order to assess emotions, lexicons are used where each of the records is somehow marked with proper emotional state or a classifier is trained with a database of well annotated data stating the nature of emotions. It is not standardized what emotions are used and how they are treated. There are many different approaches to define and classify emotions. This clause is dedicated to the description of available models for emotion classification used in related works.

4.5.2 Categorical emotion classification

Categories which are used in the research of textual emotional recognition are based on the idea of discrete emotion theory. This theory argues that there is only a limited set of emotions which a person can experience. To this category belong the following:

- Positive arousal or negative arousal.
- Discrete emotional models [i.2] and [i.31] to [i.33].

The simplest categorical classification is to decide whether the emotion is a positive or negative one. Some of the works have a similar idea, but use "happy" and "sad" state instead, as can be seen for example in the work of J.T. Hancock [i.34]. They have based their emotional recognition on leveraging the Social Information Processing model (SIP).

Another example of use are lexicons or ontologies, such as SentiWordNet of WordNet-Affect®. To annotate the records with the emotional information, very simple categorical labels were used: positive, negative, objective for SentiWordNet. Labels positive, negative, ambiguous and neutral were used for WordNet-Affect®.

Many works are building up on a research done by professor Ekman. His work is based on recognizing emotions from facial expressions. They conducted a cross-cultural study concluding that in case of facial expressions there are six basic emotions expressed the same way regardless of the subject place of origin [i.1] and [i.2]. Emotions recognized by Ekman are:

- Happiness.
- Sadness.
- Anger.
- Fear.
- Disgust.
- Surprise.

For the purposes of textual emotion recognition, there is also a neutral state considered as one of the categories to reflect neutral statements, such as headlines of selected newspapers. Some of the models based on Ekman's basic six also distinguish between positive and negative surprise [i.26].

Later on professor Ekman extended the basic set of emotions by amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame. If these emotions are expressed by face, it is difficult to distinguish between them.

A similar approach was taken by professor Izard from the University of Delaware in [i.33] where he considered 12 basic emotions: interest, joy, surprise, sadness, anger, disgust, contempt, self-hostility, fear, shame, shyness and guilt.

For the purpose of textual emotion recognition, the six basic Ekman's emotions have been used in several studies [i.5], [i.6], [i.8], [i.9], [i.11] to [i.14] and [i.27]. The other group has been used in smaller number of papers [i.15], [i.17] to [i.19] and [i.35]. It is not possible to cover all the categories, which are used by researchers, but in order to provide an overview of how many categorical emotional models are there, some of them are summarized in table 3.

Table 3: Number of emotions of emotional models

Author	# of Emotions	Emotions
Ekman	6	anger, disgust, fear, joy, sadness and surprise
Parrot	6	anger, fear, joy, love, sadness and surprise
Frijda	6	desire, happiness, interest, surprise, wonder and sorrow
Plutchik	8	acceptance, anger, anticipation, disgust, joy, fear, sadness and surprise
Tomkins	9	anger, interest, contempt, disgust, distress, fear, joy, shame and surprise
Izard	12	interest, joy, surprise, sadness, anger, disgust, contempt, self-hostility, fear, shame, shyness and guilt
Extended Ekman	18	anger, disgust, fear, joy, sadness, surprise**,** amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame
Matsumoto	22	joy, anticipation, anger, disgust, sadness, surprise, fear, acceptance, shy, pride, appreciate, calmness, admire, contempt, love, happiness, exciting, regret, ease, discomfort, respect and like

Discrete emotional categories are very popular among researchers and were used in many studies. However there are some drawbacks. Simple discrete categories are unable to reflect the valence or arousal of a particular emotion, unless somehow modified. The labels given to emotions may have a slightly different meaning when translated into other language, hence reducing its information value. As a result, a new way of assessing emotions was created and is discussed in clause 4.5.3.

4.5.3 Dimensional emotional classification

Dimensional classification methods introduce a different concept. Emotions are presented as a point or a region within a two-dimensional or multi-dimensional space. Therefore, they are not subject of assignment to one category, but to many variables. These methods extend the categorical classification by adding a scale to each of the emotions. With emotional scales for each category, the final classification forms a vector, where each item corresponds to a certain emotion and each value corresponds to emotional intensity.

Dimensionality can be applied also to categorical methods by assigning a scale to each of the categories. In this way a vector is created which represents a mixture of emotions with their respective powers. Another example is that instead of using positive or negative emotional categories, a single emotion classification with the range from -1 to 1 is used. Less than zero refers to negative emotions and above zero refers to positive emotions. The value expresses the power of the emotion.

There are many approaches which can be taken. Some special approaches were also developed to suit the needs of affective computing. The ones which will be discussed in the present document and which were recognized as the most frequent are:

- PANA Model - positive activation, negative activation.
- Circumplex model (Valence, Arousal, Power).
- Vector model.
- Plutchik emotional model.
- PAD emotional model.
- Löveheim cube of emotion.
- Hourglass model.

The PANA model is based on a single scale. It differs between negative and positive activation. Approaches like this can be used mainly for document assessment to determine agreement or disagreement.

The Circumplex model or so called Valence, Arousal, Power (VAP) model is based on a two dimensional plane with the x-axis being valence (left half-plane being unpleasant, right half-plane being pleasant) and the y-axis being arousal (upper half-plane being activation, lower half-plane being deactivation), see figure 2. The power refers to an intensity of emotional experience. For purposes of textual emotion recognition it was used in [i.36] to [i.38].

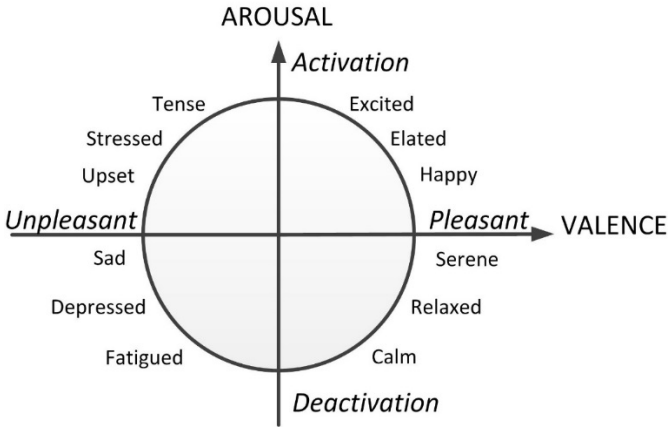


Figure 2: Circumplex model

The Vector model was used to assess drawings and words emotional perception. The emotion was expressed with two vectors in a boomerang shape. The variables used to change vectors were valence and arousal. This approach is described in detail in [i.38].

Plutchik introduces a model with 8 basic emotions organized in a circle, see figure 3, which is divided into eight separate sectors. Plutchik understands emotions as a product of adaptive biological processes. The model works with 8 basic emotions which are used to derive additional emotions by combining them. Emotions are spread in a circular plane (there exists a 3D representation as well) where the meeting point in the middle is considered neutral. The closer the emotion is to the centre, the greater its arousal.

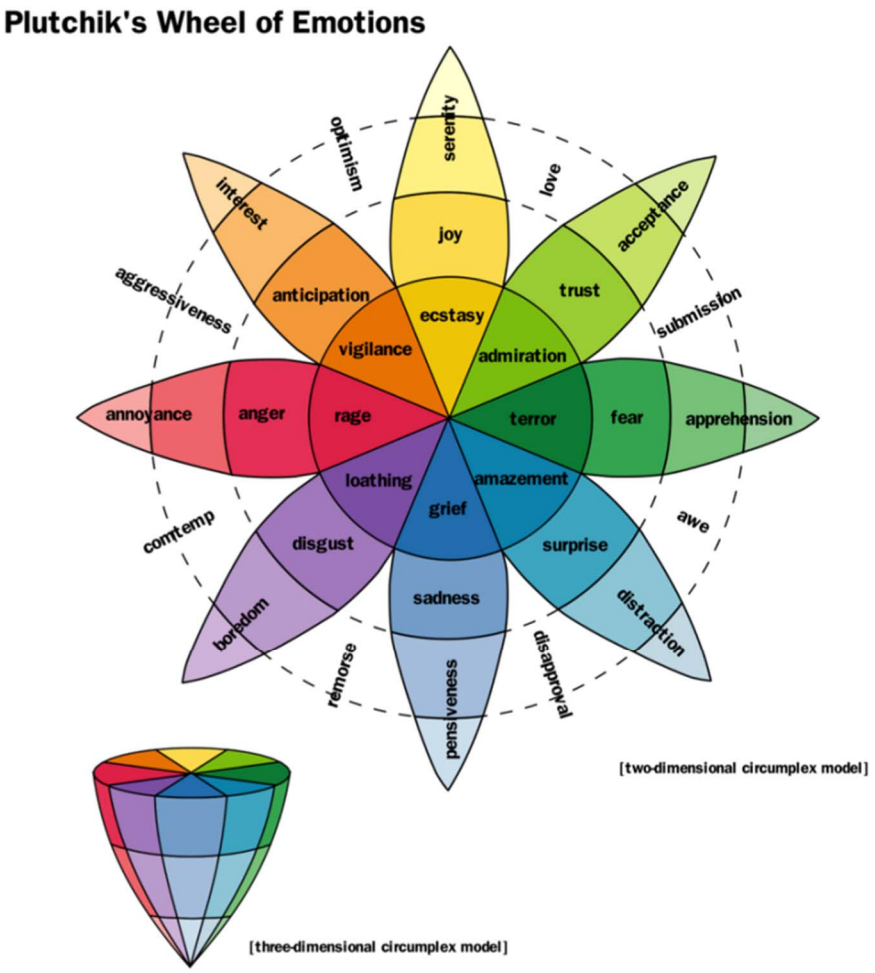
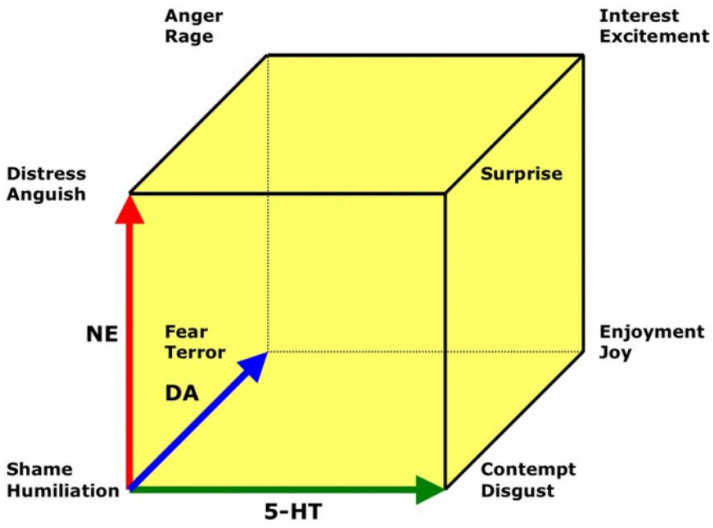


Figure 3: Plutchik's wheel of emotions

The PAD emotional model uses three scales to determine emotional scale, pleasure-displeasure, arousal-non-arousal and dominance-submissiveness.

The Löveheim cube of emotions in figure 4 is based on a different idea. It uses hormone levels to determine the emotion. Hormones considered in the cube of emotions are serotonin, dopamine and noradrenaline. It is believed, that these are the ones most responsible for experiencing emotions.



NOTE: DA dopamine;
NE noradrenaline;
5-HT serotonin.

Figure 4: The cube of emotion

Löveheim model is not used for computer interaction, but it is important to point out that emotions in terms of categories are possible to be uniquely classified when biological values are measured.

The hourglass model is a newest from the family of models presented here. It tries to overcome some issues of both categorical and dimensional models, but it is listed as one of the dimensional, since it leverages more on the idea of emotional models. The hourglass model is a biologically-inspired and psychologically-motivated approach. It means that the emotional state comes from a separate activation of a particular part of the brain [i.35] and [i.39]. There is a sentic vector, which is a 4-dimensional float describing the emotion of the subject. Emotions within this model are located in a specially shaped space which reminds the shape of a bell. The surface areas are divided into several regions, where each one of them has a specific emotional label. This allows to use categorical approach as well.

The above mentioned models of emotions are the most common. In case of computer interaction and especially of textual affect recognition, only some are used directly. The majority of published scientific papers use categorical models with scale, producing floating point vectors describing the emotion. Generally there are fewer emotional categories. Emotional recognition is very challenging since the knowledge from text is very limited, can be extremely ambiguous and depends on context.

It is important to note, that there are also other approaches to classifying emotions. But due to their use in papers the models such as Ortony, Clore and Collins model (OCC) and Five Factors Model (FFM) are not addressed here.

4.6 Algorithms

4.6.1 Introduction

Basically, emotion-detection approaches can be classified into 3 following types:

- keyword based/lexical-based;
- learning based;
- hybrid.

4.6.2 Keyword based approach

4.6.2.1 Description

In this observation the keyword based approach was the most used technique. The idea of this approach is to detect emotions in the text at the basic word level and it is very effective mostly for analysing emotion-bearing words and for simple sentences with clearly expressed emotions (detailed information about inputs in clause 4.3). As an emotion bearing words adjectives, verbs, adverbs and nouns were considered due to their ability to express emotions such as joy, anger, fear and so on [i.40].

4.6.2.2 Advantages

The main distinguishing factors of this type of approach are:

- no need of machine-learning process;
- easily implementable.

4.6.2.3 Disadvantages

However, this approach has very significant limitations:

- First of all, it has a weakness regarding the recognition of negative sentences and sentences with double negation. For example, the algorithm will have problems to determine the polarity of the sentence "The new episode of Star Wars was expected to be the exhilarating continuation of the legendary saga, but it could be way better than it is".
- The second limitation is that this approach is not able to detect emotions without obvious affect words. For example, a quote of the famous comedian John Carlin "Weather forecast for tonight: dark." would be considered as neutral. However human judgement can definitely determine the sarcastic emotion in it.

4.6.2.4 Implementation

It has been stated that the most of the keyword based approaches have similar constructions of implementation. For this method the following sequence of steps is recommended in order to perform a high-accuracy emotion detection [i.41]:

- Lexicon of emotion expressing words has to be created. For this purpose linguistic resources such as WordNet-Affect®, SentiWordNet and so on can be used (detailed description in clause 4.4).
- Text tokenization process should be applied to split the input text into parts (words, phrases, symbols, etc.).
- Obvious emotion-containing words should be detected and corresponded to the specific emotion category of the created lexicon.
- The imported text should be analysed for sentiment intensity (i.e. words "very", "too", "extremely", etc.)
- Negation check process should be implemented to detect opposite sentiment orientations (e.g. words with "not") [i.42]. For better results it is recommended to indicate negations not only in adverbs and adjectives, but in all words in the text.
- Finally, emotion is measured and extracted to "Output" box (detailed information about outputs is in clause 4.7).

Figure 5 illustrates the construction of the steps of keyword based algorithm's implementation.

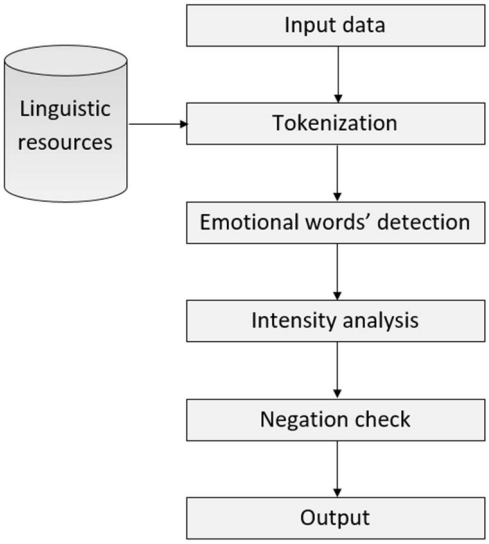


Figure 5: The construction of the keyword based approach

4.6.2.5 Related works and results

Two types of results are presented in tables of related works.

The first type is F-score which is calculated using 2 parameters called precision and recall. Precision and recall are measured using 4 parameters:

- *tp* - the percentage of true positive detected samples;
- *tn* - the percentage of true negative detected samples;
- *fp* - the percentage of false positive detected samples;
- *fn* - the percentage of false negative detected samples.

Precision is the percentage of true positive (negative) samples divided by the percentage of predicted positive (negative) samples. The unit of Precision is percent:

$$Precision(POS) = \frac{tp}{tp+fp} \times 100\ \% \quad (4.1)$$

$$Precision(NEG) = \frac{tn}{tn+fn} \times 100\ \% \quad (4.2)$$

Recall is the percentage of true positive (negative) samples divided by the percentage of actual positive (negative) samples. The unit of Recall is percent:

$$Recall(POS) = \frac{tp}{tp+fn} \times 100\ \% \quad (4.3)$$

$$Recall(NEG) = \frac{tn}{tn+fp} \times 100\ \% \quad (4.4)$$

The harmonic mean of precision and recall is called F-score [i.52]. The unit of F-score is percent:

$$F = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}\ \% \quad (4.5)$$

which is the same as:

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}\ \% \quad (4.6)$$

The second type of result is called Accuracy. This parameter presents the level of accuracy of specific algorithms and is usually defined as:

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \times 100\,\% \quad (4.7)$$

It is important to mention that the results in Accuracy column are not comparable as every algorithm designer calculated them in a slightly different way. In some cases, every emotion accuracy was counted but there was no overall accuracy of the algorithm provided. In this overview, simple average of all emotions accuracies was counted using equation (4.8):

$$Average\ Accuracy = \frac{1}{n} \cdot \sum_{i=1}^{n} Emotion_i\ Accuracy\ (\%) \quad (4.8)$$

where n - the number of calculated emotions:

- $Emotion_i$ - the value of each individual emotion accuracy.
- The unit of measurement of Accuracy is percent.

Table 4 shows works related to keyword-based approach.

Table 4: Related works using keyword based method and obtained results

Reference	Year	Algorithm	Language	F-score, %	Accuracy, %
[i.14]	2005	Keyword spotting method with syntactical sentence-level processing	English	N/A	N/A
[i.41]	2006	Keyword spotting method and semantic network	English	N/A	N/A

4.6.3 Learning based approaches

4.6.3.1 Descriptions

Learning based approaches generally have a training set which is used to train the classifier. Then the classifier is used to determine either emotion directly for a word or a more complex structure of classifiers. There are many types of learning based approaches. From the analysed papers it has been stated that the following ones were used more frequently and have higher accuracy level.

4.6.3.2 Support Vector Machines

4.6.3.2.1 Description

A commonly used and highly effective approach for identification of emotional states is the Support Vector Machines (SVM) binary classification technique, which is considered as a state of the art method. The idea of this learning based algorithm is that it uses training examples, which contain information about their category (black or white circles) (see figure 6). Successively, it creates a model which linearly classifies input data in already existing categories [i.5]. As figure 6 shows, line H_3 separates the classes with the maximum margin. Moreover, this algorithm has a special feature called kernel, which maps their inputs into high-dimensional feature spaces.

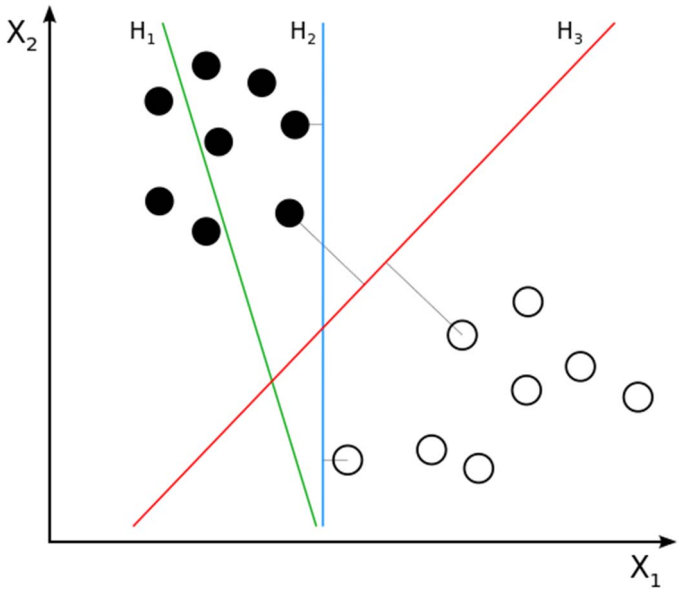


Figure 6: Basic graph of the Support Vector Machine algorithm

4.6.3.2.2 Advantages

The most important advantages worth to mentioning are:

- Ability of SVM to recognize hand-written text.
- High accuracy of prediction.
- Correct performance even when training samples are not fully accurate.

4.6.3.2.3 Disadvantages

- The SVM is able to classify data only into 2 categories. However, implementing improvement algorithms can reduce the problem of multitasking [i.6].
- It is required that Input data are fully labelled, otherwise a helping algorithm called Support Vector Clustering (SVC) should be implemented [i.43].
- Compared to other algorithms which use the same software and hardware conditions, SVM can have extremely slow speed in the test phase.

4.6.3.2.4 Related works and results

Table 5 represents some related works using SVM approach and obtained results.

Table 5: Related works using SVM algorithm and obtained results

Reference	Year	Algorithm	Language	F-score in %	Accuracy in %
[i.42]	2002	SVM with features based on unigrams	English	N/A	82,9
[i.11]	2007	SVM with General Inquirer, WordNet Affect and other features	English	N/A	73,89
[i.45]	2009	SVM with Information Gain feature extraction	Chinese	92,86 (POS) 88,28 (NEG)	91,15
[i.44]	2011	SVM with a linear kernel	Czech	N/A	80,29
[i.6]	2015	Speaker-dependent SVM with thresholding fusion	English	N/A	75,67

4.6.3.3 Naïve Bayes Classifier

4.6.3.3.1 Description

Another effective and simple learning based approach is the Naïve Bayes Classifier (NBC) algorithm. There are various types of NBC depending on representation of input text. However, the multinomial Naïve Bayes model is more accurate than other event models [i.46]. In this model, the frequencies of specific words' occurrences have been counted and demonstrated as a vector [i.47].

4.6.3.3.2 Advantages

- A very meaningful advantage of NBC is that it can be trained using only a small amount of training data.
- Has minimal consumption of CPU and memory [i.48].
- Fast training and classifying.
- Can deal with both real and discrete data.
- Inappropriate features included in training data do not have impact on its performance [i.49].

4.6.3.3.3 Disadvantages

This classifier's disadvantage is that:

- Despite the statement that features in class are independent of each other, its independence assumptions are very naïve, which means that among variables typically exist some dependencies.

4.6.3.3.4 Related works and results

Table 6 includes works related to the NBC and obtained results.

Table 6: Works related to Naïve Bayes Classifier and obtained results

Reference	Year	Algorithm	Language	F-score in %	Accuracy in %
[i.42]	2002	NBC with features based on unigrams	English	N/A	78,7
[i.51]	2012	NBC and Naïve Search	Arabic	N/A	about 85
[i.50]	2013	NBC with Facebook® Query Language query	English	72	N/A
[i.13]	2014	ERR-based NBC	English	84	N/A
[i.52]	2014	Multinomial NBC with features	Japanese	91,2	N/A

4.6.3.4 Hidden Markov Model

4.6.3.4.1 Description

The Hidden Markov Model (HMM) is also simple and frequently used learning based algorithm for emotion detection. Basically, HMM is a technique which is able to distribute classes over sequence of observations. In other words, it is possible to anticipate the future of the process as the process encapsulates all the important information concerning the past [i.53]. Figure 7 represents the graphical model of the HMM. In the model the X variables (X_1, X_2… X_T) are observed variables and the Z variables (Z_1, Z_2… Z_T) are hidden variables.

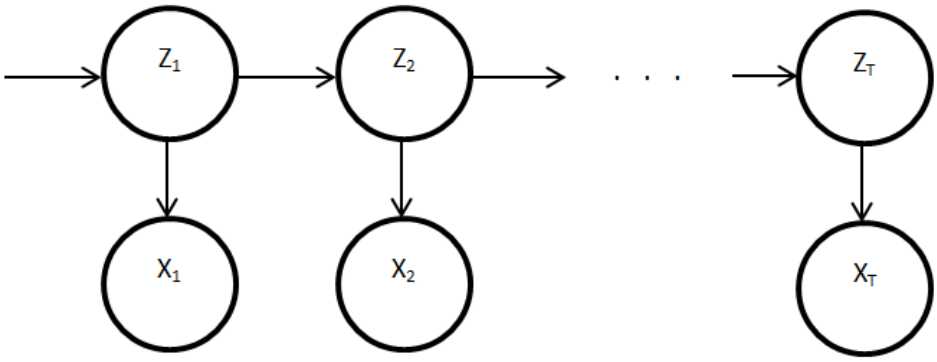


Figure 7: Graphical model of the HMM

4.6.3.4.2 Advantages

It is needed to observe only the previous state of the process.

4.6.3.4.3 Disadvantages

The main disadvantage is that it is very hard and inefficient to interpret the HMM model with big number of states and objects and over-fitting can occur. For this purpose, extensions such as Factorial HMMs, Tree structured HMMs and Switching State space models can be implemented.

4.6.3.4.4 Related works and results

Table 7 includes works related to HMM algorithm and obtained results.

Table 7: Works related to Hidden Markov Model and obtained results

Reference	Year	Algorithm	Language	F-score in %	Accuracy in %
[i.54]	2003	Continuous HMM	German and English	N/A	77,8
[i.15]	2012	A High-order HMM with Viterbi algorithm	English	35,3	N/A
[i.55]	2013	3 states HMM	English	N/A	82,95

4.6.4 Hybrid approaches

4.6.4.1 Description

This type of approaches is combination of already mentioned approach types. It is able to detect emotions based on detected keywords, learned patterns and other additional information from various dictionaries and thesauri [i.56]. Another advantage is that this approach can minimize different resources' integration complexities [i.5]. This technique is usually used when the results obtained individually by the above mentioned methods are not accurate enough. Examples of related works are listed in table 8.

4.6.4.2 Related works and results

Table 8: Works related to hybrid approaches and obtained results

Reference	Year	Algorithm	Language	F-score in %	Accuracy in %
[i.57]	2004	Hybrid SVM (PMI/Osgood and Lemmas), 100 folds	English	N/A	89
[i.9]	2013	Keyword-spotting method and rule-based method	English	76.97	N/A
[i.58]	2013	Multinomial NBC with greedy search	English	N/A	85
[i.59]	2014	NBC and SVM using Information Gain and Chi-Square methods	Chinese	N/A	71
[i.60]	2015	SVM and CRF with applied rules	Malayalam	N/A	91

4.7 Output

4.7.1 Output description

The output block is responsible for post processing of data produced by the algorithm and delivery of results to the user. The post processing is done in terms of organizing the information in a human or machine readable form, which can be later saved to a file. Delivery to the user means generating the output file and eventually visualizing the output to the user. The form of the output depends on who the user is. For social marketing it is important to know who spreads negative opinions and why. Here the emotion detector can collect negative statements for further analysis. Autistic people have trouble recognizing feelings of others in a real time. To them, a tool with real time emotion recognition in a chat can be a great help to provide them with a feedback of how their conversational partner feels. For journalists, politicians and general public, it may be interesting to use a focused search where they can look only for certain articles or webpages, for example a search of articles in favour of or against nuclear power plants (for example in search for arguments). Yet another example is search for negative reviews about a certain product to avoid a bad buy.

The form of the output is strongly tied to the emotional model. Firstly, the categorical models should be considered. With such models the word is classified and as a result assigned to a single category. A categorical output is often used in a model with two categories only (positive and negative, in favour and against). In the work of Lin [i.17], the text is processed in terms of point of view and the political articles are assigned to either Palestinian or Israeli category. A similar approach is taken for other political or review documents, where the result is in favour of or against the topic described [i.61]. Two categories are also easily used for determining whether the post on a social network has positive or negative connotation [i.50]. On a word level it can be determined whether the word has positive or negative meaning. Although two categories may be the most popular for their simplicity and straightforwardness, some models employ more categories. The six Ekman's basic emotions seem to be the most prevalent [i.5], [i.6] and [i.11] to [i.14], but there are also other categories used, for example, in education [i.15], [i.18] and [i.19].

Ekman's basic emotions overlap with dimensional classification. In dimensional classification, emotions are provided displayed as a vector (see equation (4.9) and the following two examples), xml tagged document or in another way. The output can also provide information about how confident the system is about the result in each case.

The output presented as a vector can be described as:

$$\boldsymbol{E(w) = \langle e_1, e_2, \cdots, e_N \rangle, e_n \in \langle 0; 1 \rangle, n = 1 \cdots N,} \quad (4.9)$$

where:

- w is a word, sentence or larger piece of text;
- N is the number of emotions taken into account;
- e_n is a particular emotion according to emotion model;
- E is an emotional vector.

Example of an emotional vector using Ekman's six basic emotions would then look like:

$$E(w) = \langle happiness, sadness, anger, fear, disgust, surprise \rangle$$

$$E(happy) = \langle 1, 0;\ 0, 0;\ 0, 0;\ 0, 0;\ 0, 0;\ 0, 0 \rangle$$

In some cases, the cumulative sum of all the emotions in the vector is normalized to be equal to one. This is not a rule and there are approaches, where total sum is greater than one for example in [i.62]. The output then can be coded into an xml file with tags representing emotions where necessary. A dedicated viewer can then colour the words for easy understanding. An annotated document was used, for example, in case of [i.62] and a part of a Ren-CECps database is shown on figure 8 to demonstrate possible document structure.

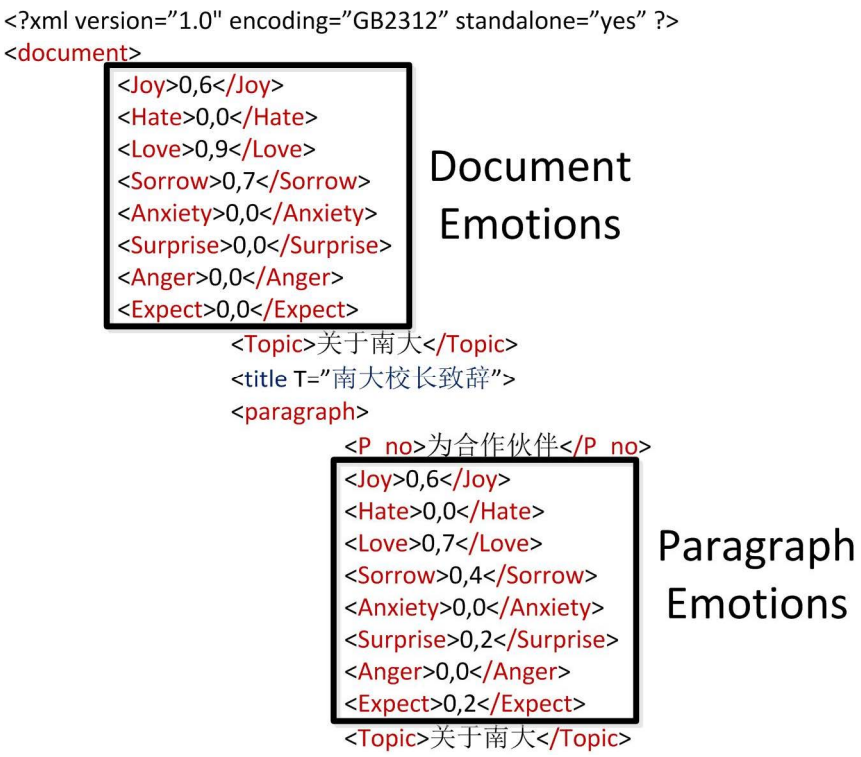


Figure 8: Emotionally tagged document

Another example of dimensional output is a sentic vector, defined in the Hourglass model [i.39] and used in [i.20] and [i.35]. Readers may encounter other types of categorical or dimensional models in the future, but the principle remains the same. An emotional detector attempts to assess the emotions in the input. Then it assigns a category or a vector based approach on a chosen model of emotions to the word or sentence according to selected granularity. The results are then saved to a separate file, or the original text is copied and accompanied with meta-data describing the emotions. Reading such files is troublesome and lengthy. Therefore, there are special means for visualizing the data based on the purpose.

4.7.2 Practical Examples

A very convenient way of expressing emotional background of words proved to be colour coding. The use of colours and other visualizations can leverage the emotional information in many different contexts. Valkanas [i.12] takes advantage of the six Ekman's basic emotions coded directly to a colour space.

Table 9: Colour coding of emotions

Emotion	Colour	Example
Neutral	White	I am Dept. of Informatics & Telecommunications (Athens, Greece)
Anger	Red	I hate it when I do something and everybody finds out! :@
Disgust	Purple	RT Retweet this if you too are offended by #ExtremeViolence #xyz
Fear	Yellow	I'm afraid this won't work out well
Joy	Green	Goaaaaaaaaaaaaaaal!!!!! Let's go@chelseafc!!! #cfc
Sadness	Blue	I miss my baby :(
Surprise	Orange	@gvalk are you serious!?

Examples of how the colour was coded to emotion and Twitter® posts which reflect the proper emotion are presented in table 9. The result implemented in a GUI are depicted in figure 9. The interface prints details related to certain events or topics and uses colour to express author's emotional state.

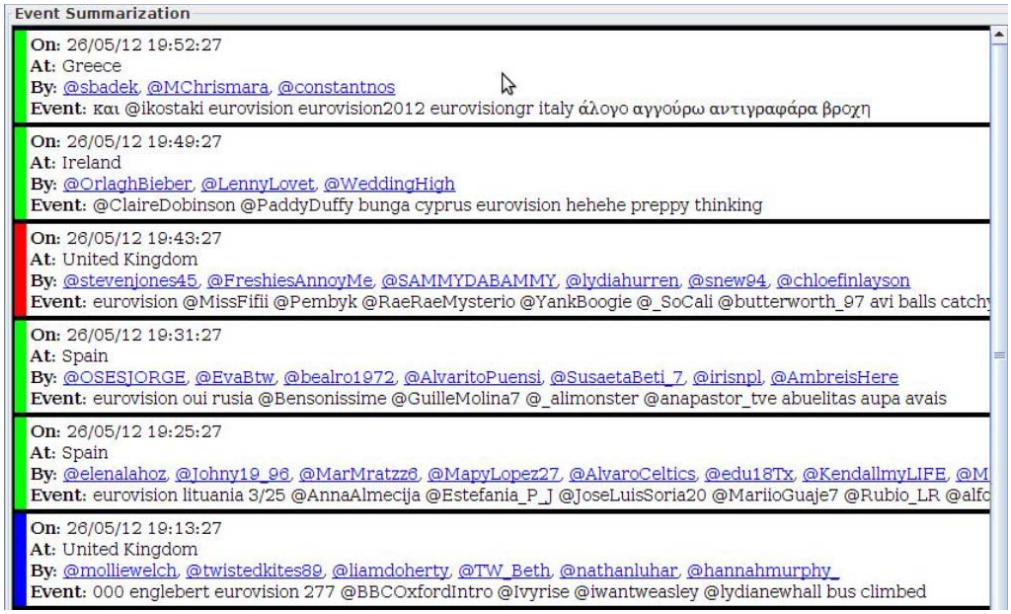


Figure 9: Twitter® with real-time emotion monitoring

This kind of system is capable of displaying emotions based on the place of origin of the tweet. Such information can be used to lay the location over a map and observe the emotional content of tweets across the world as it is depicted in figure 10.

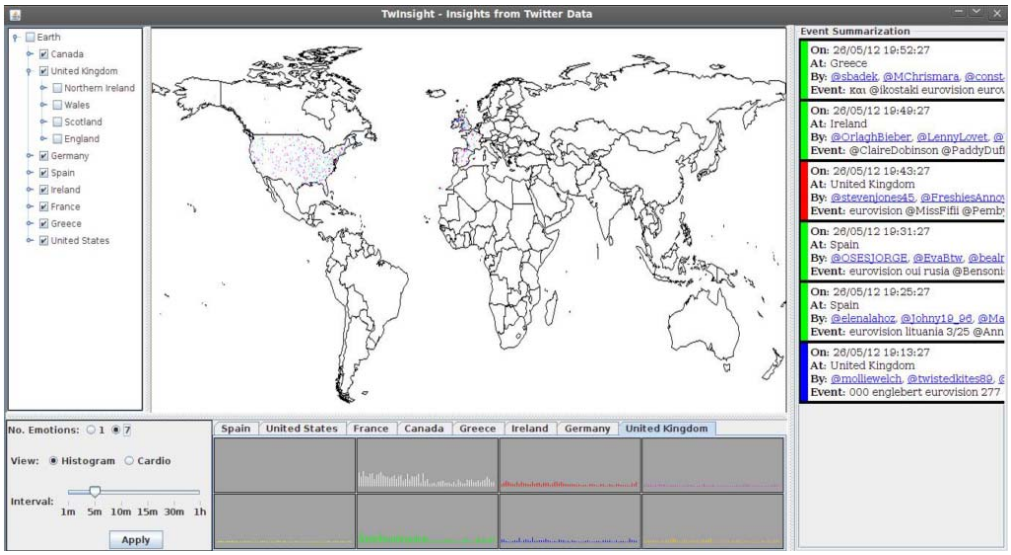


Figure 10: Twitter® emotional content displayed on the map of the world

Yet it is not the only way of applying colours to deliver emotional state. Krcadinac developed a tool in [i.9], which also works with colours, but in a different way. The program he developed is capable of real time emotion assessment and emotions are presented to the user the in form of a colourful background which changes its colour and cluster size based on the emotion in a sentence. The example displayed in figure 11 shows a sentence from a chat which was classified as a negative one and the following figure 12 shows how the background changes when the content of the sentence is evaluated as a positive one.

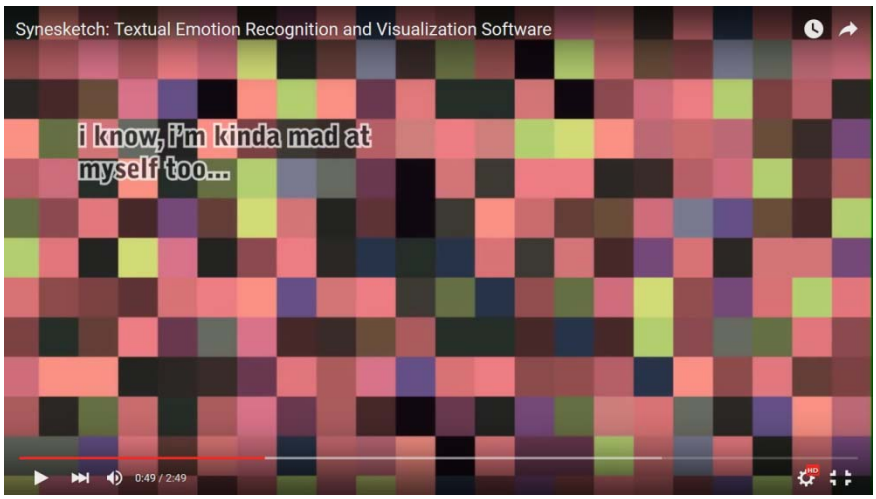


Figure 11: Example of Synesketch output for negative emotions

Darker colours represent sadder or more negative emotions, lighter and brighter colours with larger squares represent positive ones. Despite its expressive way of visualizing results it does not seem to be very practical. User would have to learn how the system operates with colours and emotions and therefore it appears more likely that this specific way of visualizing might be used for artistic interpretation of emotions.

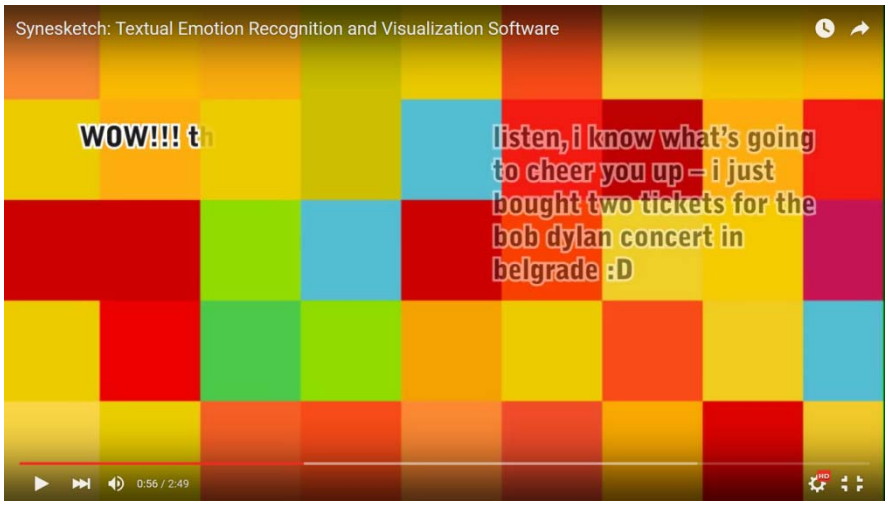


Figure 12: Synesketch positive emotions

Figure 13 and figure 14 show a different form of visualization. Instead of squares there are shapes interpreting the emotion altogether with colours. The first one works with movie reviews and the second one is a poem visualization. Although it looks compelling, for the purpose of improving user experience with respect to a chat, a different approach might be more effective.

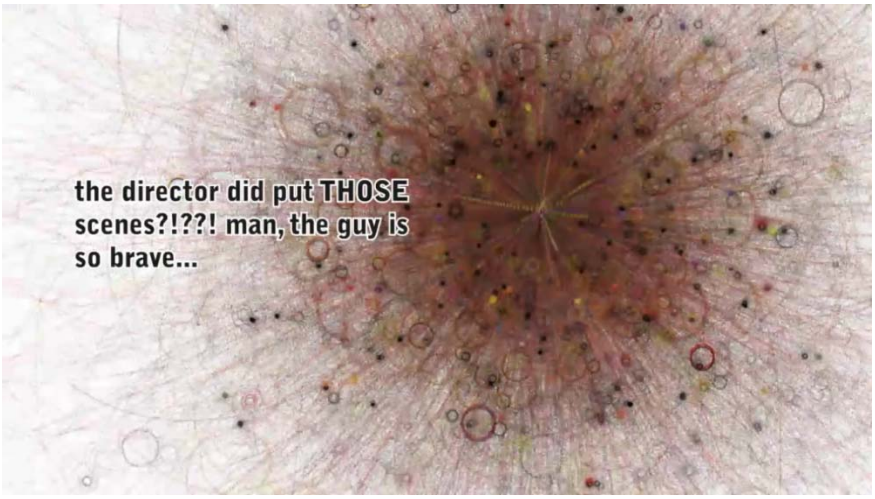


Figure 13: Synesketch reviews

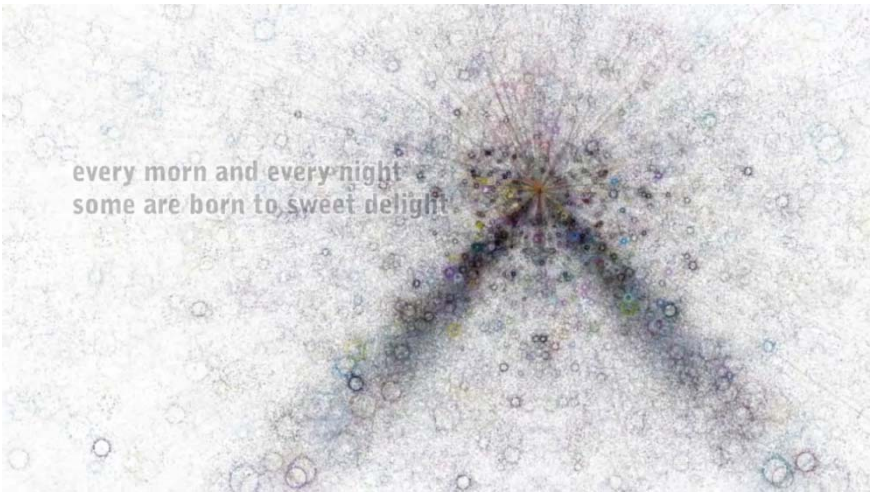


Figure 14: Poem visualization

Neviarouskaya, Prendinger and Ishizuka developed a program which displays emotions through an emotional avatar [i.27]. Its aim is to make the textual communication more human-like (see figure 4.17 of [i.27]). This application is very straightforward. The user does not need to learn any colour coding or any other way of interpreting the results. The avatar conveys emotions using facial expressions making it easy to recognize. This is a kind of tool which might help to increase the empathy among users.

In addition to providing the data to the user, the emotional output can be forwarded to another application. The aim is to provide computers with emotional feedback so they could adjust their performance of operation accordingly to optimize positive user experience. For example a tutoring program might use the emotional input to adjust its treatment of the learner to induce flow and reduce frustration and stress [i.18] and [i.19].

4.8 Final remarks on textual emotion detectors

This clause has analysed document current works, approaches and ideas in order to make understood the notion of Emotion Detector, considering all the inputs, knowledge bases, emotion models, algorithms and outputs.

Clause 4.2 introduces ways of textual emotion recognition and suggests what all the approaches have in common. As a result, figure 1 is presented describing the main parts of a general textual emotion recognition system is presented. The main important parts are linguistic resources and emotion model. Both are key to what makes emotion detection possible. Other blocks, namely input, algorithm and output are computational resources which are responsible for hard processing of the input data.

As inputs handwritten or typed text, text and audio files, as well as information imported from e-mail, social networks and instant messaging services were considered (see more in clause 4.3). After pre-processing the data is passed to the algorithm for further processing.

It has been stated in the part dealing with the algorithms (see clause 4.6) that the following three types of approach currently exist: keyword based/lexical based, learning based and hybrid. Every algorithm has naturally its advantages and limitations. Moreover, the most used and highly accurate algorithms in learning-based approaches are Support Vector Machines, Naïve Bayes classifier and Hidden Markov Model techniques. The results provided by the algorithms are then forwarded to the output block.

The output block processes the information and presents it to the user or other systems for further processing or actions.

5 Classification of emotions in scientific publications on speech recognition

5.1 Preface

Emotions frequently influence the speech behaviour. In particular, a change of prosodic parameters, i.e. fundamental frequency of human voice F_0, formants, duration of speech unit and distribution of energy in a spectrum, is a part of emotions. Emotions depend on the language carrying a message and the social background of both speakers and listeners. Current physical and mental conditions of the speaker and the listener affect emotion and consequently its influence. It is very difficult to categorize emotions, because some of them may be under certain circumstances even overlapping. Therefore, it is also very difficult to collate particular studies.

5.2 Introduction

The study and development of systems and devices that can recognize interpret, process and simulate human affects are parts of affective computing. It is an interdisciplinary field. In fact, emotions involve different components, such as subjective experience, expressive behaviour, psychophysiological changes and instrumental behaviour. Research on emotion has increased with many contributing fields including neuroscience, endocrinology, medicine and sociology. Currently, many researchers around the world study human emotions. The early research was more carried out in the field of psychology, but with the introduction of speech synthesis and automatic speaker recognition (ASR), the more psychological branch is accompanied by a more technical application-based approach. Many researchers in the field of speech technology have during the last decade worked on different aspects of emotions carried by speech.

Emotional speech recognition can even be employed by therapists as a diagnostic tool in medicine [i.63]. In psychology, emotional speech recognition methods can cope with the bulk of enormous speech data in real-time extracting the speech characteristics that convey emotion and attitude in a systematic manner [i.64]. One of the goals is to make speech synthesis sound more natural, another goal is to be able to recognize the emotive state of a speaker in a dialogue system. The other projects are motivated by the question of how recognition of emotions carried by speech could be used for business. One potential application is the detection of the emotional state in telephone call centre conversations and providing feedback to an operator or a supervisor for monitoring purposes. Another application is sorting voice mail messages according to the emotions expressed by the caller, or e.g. air force pilots' conversations in the cockpit. Some authors describe the elicited speech as the reaction to different situations [i.65], [i.66], [i.67] and [i.68].

Emotions make the language colourful and can make meaning more elaborate. Listeners also react to speakers' emotive state and adapt their behaviour depending on what kind of emotions the speaker transmits, e.g. a listener may try to show empathy to sad people, or if the speaker hesitates then the listener may try ask the speaker for clarification [i.73].

Specific projects differ in the number and type of classified emotions, acoustic characteristics, type of classifiers and precision. It is very difficult to compare methods. The phonetics-only approach to a classification sets of emotions carried by speech is insufficient. Eckman [i.74] proposed the following basic emotions: anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment and relief. Non-basic emotions are called "higher-level" emotions. Acted speech from professionals is the most reliable for emotional speech recognition because professionals can deliver speech coloured by emotions that possess a high arousal, i.e. emotions with a great amplitude or strength.

Material from radio or television is always available [i.69]. However, such material raises copyright issues and impedes the data collection distribution. An alternative is speech from interviews with specialists, such as psychologists and scientists specialized in phonetics [i.69]. Furthermore, speech from real-life situations such as oral interviews of employees when they are examined for promotion can be also used [i.70]. Parents talking to infants, when they try to keep them away from dangerous objects can be another real-life example [i.71]. Interviews between a doctor and a patient before and after medication were used in [i.72]. A major motivation comes from the desire to develop human-machine interfaces. Research by psychologists and neuroscientist has shown that emotion is closely related to decision-making and thus, emotions plays a significant role in the rational actions of human beings.

Speech can be recorded while the subject faces a machine, e.g. during telephone calls to automatic speech recognition (ASR) call centres [i.73], or when the subjects are talking to fake-ASR machines, which are operated by a human (Wizard of OZ method, WOZ) [i.74]. Automatic dialog systems with the ability to recognize emotions can respond to callers according to the detected emotional state or they can pass control over to human operators. Automatic emotion recognizers can be viewed as systems that assign category labels to emotional states. Giving commands to a robot is another idea explored [i.75]. Speech can be also recorded during imposed stressed situations.

Emotions represent a mental state of a living organism possibly accompanied by body or face motion and by a change of physiological state of a subject. Some authors describes the elicited speech as the reaction to different situations [i.65], [i.66], [i.67] and [i.76]. Only 10 % of information is included in utterances. Especially, it is intonation which is related to the meaning of the sentence and its emotional timbre.

The first problem one faces in starting to investigate emotions in speech is the problem of choosing valid data. Generally materials from three categories are used in investigating emotional speech. They are database of staged emotions, database of emotions evoked by external stimulus and database of spontaneous emotions. The first of them is created by professional actors (it has least weight from the point of view of likelihood); the recordings are make at sound laboratories, without noise. The external stimuli are used for second type of emotions, the subject answers to questions which provoke an emotion (a short video is used); these records have a higher likelihood than the first type. The highest likelihood characterizes spontaneous emotions, but it is intrinsically very difficult construct them. The problems come from legal and ethical aspects. It is essential for the field to be well versed in the ethical issues involved in collecting emotionally coloured data. Most institutions now routinely require ethical approval for any work with humans and is not routine to give approval for a producer to take actions that involve eliciting negative emotions or deception, both of which are very common in emotion elicitation scenarios. Therefore records from TV and radio or a conversation at pilot's cockpits are exploited.

Simulated emotions create corpuses in most cases. [i.77] contains examples of speech material from which corpuses are created . Emotional speech databases are recorded for different languages (12 languages for databases described in the present document). Most databases are for English (10) and German (8). The other variants are composed for Japanese (3), Dutch (2), Spanish (2), in singles language for Danish, Hebrew, Sweden, Russian, Chinese, Czech and multilingual (Slovenian, French, English and Spanish). The number of emotions varies between two and eight. The most frequent emotion is anger, followed in descending order by sadness, fear, happy, neutral, disgust, joy, surprise, boredom, stress, anxiety, despair, panic and shame. Also the number of subjects (speakers) is different - from one subject to 780 subjects. Most often they are actors. The other speakers are volunteers, TV newsreaders, students, soldiers, drivers, pilots and passengers. The speaker is different too (female, male and children). The numbers of databases available for each emotion are summarized in table 10.

Table 10: Emotions recorded in databases

Emotions	Number of databases
Anger	26
Sadness	22
Happiness	13
Fear	14
Neutral	12
Disgust	10
Joy	9
Surprise	6
Boredom	5
Contempt	2
Dissatisfaction	2
Shame, pride, worry, despair, humour, etc.	1

5.3 Basic information about speech emotions

Robert Plutchik defines emotions as an inferred complex sequence of reactions to a stimulus including cognitive evaluations, subjective changes, autonomic and neural arousal, impulses to action and behaviour designed to have an effect upon the stimulus that initiated the complex sequence. Four major traditions that have developed in the historical literature on emotions are discussed: Darwin's evolutionary tradition, the psychophysiological theory of William James, the neurological theory of Walter Cannon and the dynamic tradition of Freud. Eight basic reactions are proposed because they are seen as the prototypes of all emotions: rage, loathing, grief, terror, adoration, amazement, ecstasy and vigilance [i.78].

Klaus Scherer from University of Geneva published one of the basic articles in emotions area. He reviews the evidence on listeners' ability to accurately identify a speaker emotion from voice cues alone, the research efforts trying to isolate the acoustic features that determine listener judgments and the findings on actual acoustic concomitants of a speaker emotional state. Finally, based on speculations about the joint origin of speech and vocal music in non-linguistic affect vocalizations, similarities of emotion expression in speech and music were discussed [i.79].

Greek researchers from Centre for Research and Technology Hellas in [i.80] focus on emotional speech recognition aims to automatically classify speech units (e.g. utterances) into emotional states, such as anger, happiness, neutral (low scores on all emotions indicate that an affecting stimulus is absent), sadness and surprise. The major contribution of this paper is to rate the discriminating capability of a set of features for emotional speech classification when gender information is taken into consideration. A total of 87 features has been calculated over 500 utterances of the Danish Emotional Speech database. Data from this database was tested by the Bayes classifier and classification success was 54 % (for both genders), 61 % for males and 57 % for females with the classification of the five classes (for anger, pleasure, sadness, surprise and the neutral state (as explained above)). Also the focus was on providing an overview of emotional speech classification having in mind three goals in [i.81]. The first goal was to provide an up-to-date record of the available emotional speech data collections. The number of emotional states, the language, the number of speakers and the kind of speech were briefly addressed. The second goal was to present the most frequent acoustic features used for emotional speech classification and to assess how the emotion affects them. Typical features are the pitch, the formants, the vocal tract cross-section areas, the mel-frequency cepstral coefficients, the Teager energy operator-based features (processing the speech signal in noisy environment), the intensity of the speech signal and the speech rate. The third goal was to review appropriate techniques in order to classify speech into emotional states.

Valery Petrushin from Oakland University in [i.82] describes two experimental studies on vocal emotion expression and recognition. The first study deals with a corpus of 700 short utterances expressing five emotions: happiness, anger, sadness, fear and normal state, which were portrayed by 30 non-professional actors. After evaluation a part of this corpus was used to extract features and train a multilayer neural network with backpropagation training algorithm. Some statistics of the pitch, the first and second formants, energy and speaking rate were selected as relevant features using feature selection techniques. The accuracy for normal state was 60 % to 75 %, for happiness 60 % to 70 %, for anger 70 % to 80 %, for sadness 70 % to 85 % and for fear 35 % to 55 %. The total average accuracy was about 70 %. The second study uses a corpus of 56 telephone messages of varying length (from 15 seconds to 90 seconds) expressing mostly normal and angry emotions that were recorded by eighteen non-professional actors. These utterances were used for creating recognizers by means of the methodology developed in the first study.

The recognizers were able to distinguish between two states: "agitation" which includes anger, happiness and fear and "calm" which includes normal state and sadness with average accuracy of 77 %.

The classification of three emotions (sadness, anger and neutral state (as explained above)) for human-computer communication is described in [i.83]. Authors focused on the emotional state of a speaker and showed how acoustic and prosodic information can be used to detect the underlying emotional state of a speaker. The hidden Markov Model architecture was applied. The corpus contained 50 sentences (questions, statements and orders). The sentence length varied from two to 12 words. The corpus was comprised of 291 word tokens (87 types). Five students of dramatic school pronounced the sentence according to emotional label (for happy, sad, angry and afraid). Also 50 neutral sentences were pronounced.

Researchers showed that music and speech are partly processed by common parts of the brain. Therefore they have something in common. The people who lack musicality (they have a music perception deficit) have problems with emotion recognition, which is voiced by spoken words. 4 % persons of the human population belong to such group. In speech, it is possible to find a parameter which corresponds to the tone relationship and for speech emotions it may have perceptional important values even in changes of frequency range appearing in speech intonation. A compound tone is a tone of all music instruments but also a tone of speech: the spectrum of these tones is a set of integral multiples of a fundamental frequency. This set is called harmonic row or series. It should also be noted that from the listener's standpoint it is perceived as the only one tone. The amplitude of tones in the harmonic row diminishes with increasing index. At that point when the amplitude of fundamental tones is zero or lower, it is still unambiguously perceived as identical (e.g. narrowband voice, fundamental tone, which is responsible for perception of intonation, is unfiltered, but one can still perceive intonation) [i.84] and [i.85]. A description of the five emotional states (pleasure, sadness, fear, anger and neutral state (as explained above)) is undertaken in [i.86]. Emotional speech synthesis is based on rules that describe the behaviour of the pitch frequency along time to generate time-pitch values. Pitch values fluctuate within a certain range depending on the intended emotion. These studies have shown that a slight 4 Hz modification of pitch frequency is sufficient to make a significant change in the emotional state of speech. This approach relies on mapping the pitch frequency values to the 12 semitone melodic scale and extracting semitonic intervals for each emotional state. The authors use PRAAT analysis tool.

A group of researchers active in the medical domain undertook an emotion project [i.87]. The authors have studied the influence of vocal emotions on speech understanding. Seven emotions (anger, disgust, fear, sadness, neutral (as explained above), happiness and pleasant surprise) were tested with younger and older listeners. Emotion was the within-subjects variable: all participants heard speech stimuli consisting of a carrier phrase followed by a target word spoken by either a younger or an older talker, with an equal number of stimuli portraying each of these seven vocal emotions. Listener age (younger, older), condition (mixed, blocked) and talker age (younger, older) were the main inter-subjects variables. 56 students (mean age = 18,3 years) were recruited from an undergraduate psychology course; 56 older adults (mean age = 72,3 years) were recruited from a volunteer pool. All participants had clinically normal pure tone audiometric thresholds.

It is felt correct to state that three areas of emotions research are found in publications. They are emotional speech synthesis, recognition (or classification) of emotions and deployment of agents for decoding and expressing emotions.

5.4 Language

Aneta Pavlenko from TESOL at the College of Education, Temple University, Philadelphia is the author of many publications about emotions and bilingualism resp. multilingualism [i.88]. In bilingual societies, people frequently switch between languages, particularly in emotional situations. Researchers from the University of California, Berkeley and Bard College say that such code-switching is a strategy to express emotions differently, particularly when communicating with children. Bilingual parents may use a specific language to express an emotional concept because they feel that the corresponding language provides a better cultural context for expressing the particular emotion [i.89]. The first large-scale investigation on how multilinguals feel about their languages and use them to communicate emotion is goal a book written by Jean-Marc Dewaele [i.90]. Using a combination of quantitative and qualitative approaches, Jean-Marc Dewaele looks at the factors that affect multilinguals' self-perceived competence, attitudes, communicative anxiety, language choice and code-switching when expressing feelings, anger and when swearing. Nearly 1 600 multilinguals from all over the world participated in the research. The results suggest that how and when a language was learnt determines future use and communicative anxiety. Aspects such as present use of the language, the total number of languages known and the level of emotional intelligence also play an important role. Interviews with participants reveal the importance of cultural factors and show how the slow process of acculturation in a new community is accompanied by gradual changes in language preferences to communicate emotions.

5.5 Existing Corpora

5.5.1 Preface

The collection of emotion corpora or databases is a basic necessity. Many authors divide emotions into primary and secondary or into active and passive. Its number is different for each particular group. The following text and tables provide more detailed information about existing databases. The overview is shown in table 11.

Table 11: Language used in databases recording

Language	Number of databases	Publications
English	12	[i.91], [i.97] to [i.108] and [i.112]
German	6	[i.109] to [i.118]
Slovenian	1	[i.96] and [i.97]
Japanese	3	[i.119], [i.120] and [i.121]
Spanish	2	[i.97], [i.98], [i.124] and [i.125]
Dutch	2	[i.110], [i.126] and [i.128]
French	4	[i.97] and [i.98]
Swedish	1	[i.128]
Hebrew	1	[i.127]
Danish	1	[i.126]
Italian	1	[i.131]
Chinese	1	[i.129]
Russian	1	[i.130]
English and German	2	[i.109] and [i.111]
Multi-Language (see note)	2	[i.117]
Artificial Language	1	[i.111]
Czech	1	[i.85] and [i.133]
NOTE: Four languages: one database with French, English, Slovenian and Spanish, and one database with French, English, German and Italian.		

5.5.2 Introduction

The collection of emotional speech databases is a necessary condition for emotion speech pre-processing. A large source of emotional databases information created so far is presented in [i.77]. Authors Dimitrios Ververidis and Constantine Kotropoulos reviewed 32 emotional speech databases. Each database consists of a corpus of human speech pronounced under different emotional conditions. A basic description of each database and its applications is provided. Databases are grouped by language in English, German, Japanese, Dutch, Spanish, Danish, Hebrew, Swedish, Chinese, Russian and multilingual emotion speech databases. Databases include existing corpora of spontaneous speech (they are freely available), for example the Belfast database [i.91] - from TV programs, the Leeds-Reading Emotion in Speech Corpus (e.g. [i.92]) - from TV programs, the JST database (e.g. [i.93]) - the natural speech in natural situations, the SUSAS corpus [i.94] - which consists of air force pilots conversation in the cockpit, the EISP corpus [i.92] - the EISP corpus consists of recordings of TV and radio documentaries. In this text 36 emotional speech databases are described.

The remaining databases, listed below are freely available. Each database consists of a corpus of human speech pronounced under different emotional conditions. The databases are recorded by an overwhelming majority by native speakers. Also native listeners test emotions classification correctness. From the databases listed below 15 and 28 are the exemption (see clauses 5.5.4.6 and 5.5.10.2, respectively). The text also has its applications or application context. The conclusion of this part is that an automatic emotion recognition of emotion carried by speech can attain a correct classification maximally 50 % for the basic emotions, i.e. anger, disgust, fear, happiness, sadness and surprise according to [i.95]. The following text is adopted from [i.77].

5.5.3 English speech emotion databases

5.5.3.1 Preface

Most databases concentrate on English language (16 databases - separately or in combination with other languages).

5.5.3.2 Database 1

R. Cowie and E. Cowie [i.98], [i.99] and [i.100] constructed this database at the Queen's University of Belfast. It contains emotional speech in five emotional states: anger, sadness, happiness, fear and neutral (see clause 5.3 for details). The speakers are 40 volunteers (20 female, 20 male) aged between 18 years and 69 years. The subjects read five passages of seven to eight sentences written in an appropriate emotional tone and content for each emotional state. Each passage has strong relationship with the corresponding emotional state.

5.5.3.3 Database 2

R. Cowie and M. Schroder constructed Belfast Natural Database at Queen's University [i.91]. The database is designed to sample genuine emotional states and to allow exploration of the emotions through time. Two kinds of recordings took place. One was recorded in studio and the other from TV programs. A total of 239 clips (10 seconds to 60 seconds) is included in the database. The clip length has been selected to be quite long in order to reveal the development of emotion over time. The studio recordings consist of two parts. The first part contains conversations between students on different topics, which provoke strong feelings. The second part contains audio-visual recordings of interviews (one-to-one) involving a researcher with fieldwork experience and a number of friends.

5.5.3.4 Database 3

R. Cole and his assistants at the University of Colorado recorded Kids' Audio Speech Corpus NSF/ITR Reading Project [i.101]. The aim of the project was to collect sufficient audio and video data from kids in order to enable the development of auditory and visual recognition systems, which enable face-to-face conversational interaction with electronic teachers. The Kids' Audio speech Corpus is not clearly oriented to elicit emotions. Only 1 000 out of 45 000 utterances are emotion oriented.

5.5.3.5 Database 4

M. Liberman, Kelly Davis and Murray Grossman at the University of Pennsylvania constructed the Emotional Prosody Speech and Transcripts [i.102]. The database consists of 9 hours of speech data. It contains speech in 15 emotional categories, such as hot anger, cold anger, panic-anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt.

5.5.3.6 Database 5

J. Hansen at the University of Colorado Boulder has constructed in 1999 an emotional speech corpus the so called SUSAS (Speech Under Simulated and Actual Stress). The database contains voice from 32 speakers (13 female and 19 male) with ages ranging from 22 to 76 years old. In addition, four military helicopter pilots were recorded during the flight. Words from a vocabulary of 35 aircraft communication words make up the database. The total number of utterances is 16 000. SUSAS database is distributed by the LDC [i.102].

5.5.3.7 Database 6

C. Pereira at Macquire University constructed this database [i.103]. The database consists of 40 sentences uttered by two actors in five emotional categories. There are two repetitions of these 40 utterances, thus creating 80 presentations altogether. In the study, 31 normal hearing subjects (18 men and 13 women) rated all the utterances. The listeners rated each utterance on six Likert intensity scales [i.104].

5.5.3.8 Database 7

M. Edgington at BT Labs, UK collected this emotional speech database for training a voice synthesizer [i.105]. One professional male actor was employed. The database contains speech in six emotional categories, such as anger, fear, sadness, boredom, happiness and neutral (see clause 5.3 for details). A laryngograph was also recorded. 13 subjects identified the emotions with 79,3 % score rate. The database also includes the signal energy, syllabic duration and the fundamental frequency of each phoneme.

5.5.3.9 Database 8

T. S. Polzin and A. H. Waibel at the Carnegie Mellon University constructed this emotional speech database [i.106]. The database consists of emotional speech in five emotional categories. The corpus was comprised of 291 word tokens per emotion per speaker. The sentence length varies from 2 to 12 words. These sentences are comprised of questions, statements and orders. The database was evaluated by other people. The recognition rate of each emotion is 70 %. The baseline is 25 % for random guessing.

5.5.3.10 Database 9

V. Petrushin at the Centre for Strategic Technology Research, Accenture constructed this database in order to train neural networks in speech recognition [i.107]. It is divided into two studies. The first study deals with a corpus of 700 short utterances expressed by 30 professional actors. The corpus contains emotional speech in five emotion categories, such as happiness, anger, sadness, fear and normal, which were portrayed by thirty non-professional actors. In the second study, 56 telephone messages were recorded. The length of each message is from 15 seconds to 90 seconds.

5.5.3.11 Database 10

R. Fernandez at MIT labs constructed emotional speech this database [i.108]. The subjects are drivers who were asked to sum up two numbers while driving a car. The questions produced by a speech synthesizer and the sum of the numbers was less than 100. The two independent variables in this experiment were the driving speed and the frequency at which the driver had to solve the math questions. Every subject drove in two speed conditions, one at 60 miles per hour (approximately 96,6 km/h) and one at 120 miles per hour (approximately 193,1 km/h). In the low speed, the subject was asked for an answer every nine seconds. In the high speed, the subject was asked to sum up the numbers every four seconds.

5.5.4 German speech emotion databases

5.5.4.1 Preface

Most databases concentrate on German language (9 databases - separately or in combination with other languages).

5.5.4.2 Database 11

The Verbmobil database was recorded at the University of Hamburg [i.109] and [i.110]. It is non-neutral emotion oriented, because it contains mainly anger and dissatisfaction. The database contains voice from 58 native German speakers, between 19 and 61 years old (29 male, 29 female), while they were speaking to an emulated ASR system. From distance, a researcher ("a wizard") controlled the ASR response, in such a way that the speaker believed that he was speaking to a machine. The above dialogues are named "Wizard of OZ" dialogues. Prosodic properties of the emotion, such as syllable lengthening and word emphasis are also annotated. This database can be used for training speech recognition systems.

5.5.4.3 Database 12

The SmartKom Multimodal Corpus was constructed at the Institute for Phonetics and Oral Communication in Munich, with national funding [i.109] and [i.111]. The database consists of "Wizard of OZ" dialogues (like Verbmobil database) in German and English. The database contains multiple audio channels (which follow speakers' position changes) and two video channels (face, body from side). It is oriented to emotion as it includes mainly anger and dissatisfaction. The aim of the project was to build a gesture and voice recognition module for human-computer interfaces.

5.5.4.4 Database 13

W. F. Sendlmeier et al. at the Technical University of Berlin constructed another emotional speech database [i.112] and [i.113]. The database consists of emotional speech in seven emotion categories. Each one of the 10 professional actors expresses 10 words and five sentences in all the emotional categories. The corpus was evaluated by 25 subjects who classified each emotion with a score rate of 80 %.

5.5.4.5 Database 14

K. Alter at the Max-Planck-Institute of Cognitive Neuroscience constructed an emotional speech database for medical purposes [i.114]. Electroencephalogram (EEG) recordings were also taken. The aim of the project was to relate emotions, which are recognized from speech with a location in the human brain. A trained female fluent speaker was employed. The database contains speech in three emotional categories. Twenty subjects judged both the semantic content and the prosodic feature on a five-point scale.

5.5.4.6 Database 15

K. Scherer at the University of Geneva constructed this database [i.111]. The purpose of his study was to bring into light the differences in emotional speech perception between people from different countries. Results showed that the Indonesian people understand the emotions in speech in a different way. Four German professional actors were employed. The sentences derived from an artificial language, which was constructed by a professional phonetician.

5.5.4.7 Database 16

B. Wendt and H. Scheich at Leibniz Institute of Neurobiology constructed the Magdeburger Prosodie Korpus [i.115]. The aim was to construct a brain map of emotions. The database contains emotional speech in six emotional categories. The database contains also the word accentuation, word length, speed rate, abstraction/concreteness, categorizations and phonetic minimal pairs. The total number of utterances is 4200 nouns and pseudo words.

5.5.4.8 Database 17

M. Schroeder and M. Grice recorded a database of diphones that can be used for emotional speech synthesis [i.116]. One male speaker of standard German produced a full German diphone set for each of three degrees of vocal effort: "soft", "modal" and "loud". Four experts verified the vocal effort, the pitch constancy and the phonetic correctness.

5.5.4.9 Database 18

M. Schroder has constructed an emotional speech database, which consists of "affect bursts" [i.117]. His study shows that affect bursts, presented without context, can convey a clearly identifiable emotional meaning. Professionals selected the affect bursts from the German literature. Altogether, the database comprises about 80 different affect bursts. The database contains speech in 10 emotion categories. In order to define the intended emotions for the recordings, a frame story was constructed for each of the 10 emotions.

5.5.5 Japanese speech emotion databases

5.5.5.1 Preface

Three databases focus on Japanese language. Two of them are collected at the ATR Laboratories, one of them at the Kyoto Institute of Technology.

5.5.5.2 Database 19

R. Nakatsu et al. at the ATR Laboratories constructed an emotional speech database in Japanese [i.118]. The database contains speech in eight emotion categories. The project employed 100 native speakers (50 male and 50 female) and one professional radio speaker. The professional speaker was told to read 100 neutral words in eight emotional manners. The 100 ordinary speakers were asked to mimic the manner of the professional actor and say the same amount of words. The total amount of utterances is 80 000 words.

5.5.5.3 Database 20

Y. Niimi et al. at the Kyoto Institute of Technology, Matsugasaki, Japan developed another emotion speech database in Japanese [i.119]. It consists of VCV (Vowel Consonant Vowel) segments for each of the three emotion speech categories, such as anger, sadness and joy. These VCVs can generate any accent pattern of Japanese. They were collected from a corpus of 400 linguistic unbiased utterances. The utterances were analysed to derive a guideline for designing VCV databases and to derive an equation for each phoneme, which can predict its duration based on its surrounding phonemic and linguistic context. 12 people judged the database and they recognized each emotion with a rate of 84 %.

5.5.5.4 Database 21

Iida and N. Campbell at ATR Laboratories constructed a third emotional speech database in Japanese [i.120]. The aim of the project was emotional speech synthesis for disabled people. The database contains emotional speech in three emotion categories, such as joy, anger and sadness. The emotions are simulated but not exaggerated. The database consists of monologue texts collected from newspapers, web sites, self-published autobiographies of disabled people, essays and columns. Some expressions typical to each emotion were inserted in appropriate places in order to enhance the expression of each target emotion.

5.5.6 Dutch emotion speech databases

5.5.6.1 Preface

Two databases focus on Dutch language. It is important to note that the Groningen database is only partially oriented to emotions.

5.5.6.2 Database 22

Groningen database. It is constructed at the Psychology School at Groningen University in The Netherlands and is distributed by ELRA [i.110]. It contains 20 hours of Dutch speech. It is important to note that the database is only partially oriented to emotion. An electroglottograph and an orthographic transcription are also included. The total number of speakers is 238. They are not actors and the emotions are forced rather than naturally produced. The database consists of short texts, short sentences, digits, monosyllabic words and long vowels.

5.5.6.3 Database 23

S. Mozziconacci and his assistant collected an emotional speech database in order to study the relationship between speed in speaking and emotion [i.121] and [i.122]. The database contains emotional speech in seven emotion categories, such as joy, boredom, anger, sadness, fear, indignation and neutral (see clause 5.3 for details). The speech material used in the study consists of 315 utterances. Each of the three speakers reads five sentences, which have a semantically neutral content. Twenty-four judges evaluated the utterances and two intonation experts labelled them.

5.5.7 Spanish emotion speech databases

5.5.7.1 Preface

Two databases consist of Spanish language only (Databases 24 and 25). Additionally the Maribor database covering four languages contains Spanish language, too.

5.5.7.2 Database 24

Spanish Emotional Speech database (SES). J. M. Montero and his assistants constructed a Spanish emotional speech database in 1998 [i.123]. It contains emotional speech in four emotion categories, such as sadness, happiness, anger and neutral (see clause 5.3 for details). Fifteen subjects identified each emotion with a score of 85 %. The labelling of the database is semi-automatic. The corpus consists of three short passages (four to five sentences), 15 short sentences and 30 isolated words. All have neutral lexical, syntactical and semantical meaning.

5.5.7.3 Database 25

I. Iriondo [i.124] at the University of R. L. of Barcelona also recorded an emotional speech database. It contains emotional speech in seven emotion categories. Each of the eight actors reads two texts in three emotional intensities. The speech was rated-judged by 1 054 students during a perception test. From the 336 discourses, only 34 passed the perception test.

5.5.8 Danish emotion speech database

5.5.8.1 Preface

Four emotions and neutral (see clause 5.3 for details) speech samples are processed in one Danish speech emotion database.

5.5.8.2 Database 26

I. F. Engberg and A.V. Hansen at the Centre for Person Communication at Aalborg University recorded the Danish Emotional Speech database (DES) [i.125]. T. Brondsted wrote the phonetical transcription. The construction of DES was a part of Voice Attitudes and Emotions in Speech Synthesis (VAESS) project. The database contains emotional speech in five emotion categories, such as surprise, happiness, anger, sadness and neutral (see clause 5.3 for details). The database consists of two words (yes, no), nine sentences and two passages. Twenty judges (native speakers from 18 to 58 years old) verified the emotions with a score rate of 67 %.

5.5.9 Hebrew emotion speech database

5.5.9.1 Preface

Three medical parameters are taken into account in one Hebrew database.

5.5.9.2 Database 27

At the faculty of Holon Academic Institute of Technology at Israel, N. Amir et al. recorded emotional a multi-modal speech database [i.126]. It contains emotional speech in 5 emotion categories. The database consists of emotional speech, electromyogram of the curragator (a muscle of the upper face which assists in expressing an emotion), heart rate and galvanic resistance that is a sweat indicator. The subjects (40 students) were told to recall an emotional situation of their life and speak about that. In his study N. Amir found that there is not an absolute clear way of discovering a true emotion carried by speech.

5.5.10 Swedish emotion speech database

5.5.10.1 Preface

Speakers from four different native speaking languages contributed for the Swedish database. They are Swedish, Spanish, Finnish and English speakers.

5.5.10.2 Database 28

A. Abelin and his assistants recorded this emotional speech database [i.127]. It contains emotional speech in nine emotion categories, such as joy, surprise, sadness, fear, shyness, anger, dominance, disgust and neutral (see clause 5.3 for details). Different nationality listeners classified the emotional utterances to an emotional state. The listener group consisted of 35 native Swedish speakers, 23 native Spanish speakers, 23 native Finnish speakers and 12 native English speakers. The non-Swedish listeners were Swedish immigrants and all had knowledge of Swedish, of varying proficiency.

5.5.11 Chinese emotion speech database

5.5.11.1 Preface

Three emotions and neutral speech segments were created by three Chinese students and by avatars.

5.5.11.2 Database 29

F. Yu et al. at the Microsoft Research China recorded this emotional speech database [i.128]. It contains speech segments from Chinese teleplays in four emotion categories, such as anger, happiness, sadness and neutral (see clause 5.3 for details). Four persons tagged the 2 000 utterances. Each person tagged all the utterances. When two or more persons agreed in their tag, the utterance got their tag while all other utterances were thrown away. After tagging several times, only 721 utterances remained.

5.5.12 Russian emotion speech database

5.5.12.1 Preface

A big database was collected at the Meikai University in Japan by native Russian students for Russian language.

5.5.12.2 Database 30

RUSSian LANguage Affective speech (RUSSLANA). V. Makarova and V.A. Petrushin collected this emotional speech database at the Meikai University in Japan [i.129]. In total, 3 660 sentences uttered by 61 native Russian speakers (of which 12 males) from 16 to 28 years old are included in this database. Features of speech like energy, pitch and formants curves are also included.

5.5.13 Multilingual emotion speech database

5.5.13.1 Preface

Three Swiss official languages (French, German and Italian) and English language in database 31 and four languages (i.e. English, Slovenian, French and Spanish) in database 32 are used in the following databases.

5.5.13.2 Database 31

Lost Luggage study. K. Scherer has recorded another emotional speech database [i.117]. The recordings took place inside Geneva International Airport. The subjects are 109 airline passengers waiting in vain for their luggage to arrive on the belt.

5.5.13.3 Database 32

The database was recorded at the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia [i.96]. It contains emotional speech in six emotion categories, such as disgust, surprise, joy, fear, anger and sadness. Two neutral emotions were also included: fast loud and low soft. It is seen that the emotion categories are compliant with MPEG-4 [i.97]. Four languages (i.e. English, Slovenian, French and Spanish) were used in all speech recordings. The database contains 186 utterances per each emotion category. These utterances are divided in isolated words, sentences both affirmative and interrogative and a passage.

5.5.14 Four other databases

5.5.14.1 Preface

Italian, Czech and English languages (different accents and non-native speakers) are used in the following emotion databases.

5.5.14.2 Database 33

EMOVO - Emotional corpus applicable to Italian language. It is a database built from voices of up to six actors who played 14 sentences simulating seven emotional states (disgust, fear, anger, joy, surprise, sadness and neutral (see clause 5.3 for details)).The recordings were made with professional equipment in the Fondazione Ugo Bordoni laboratories. The validation test was successful with 80 % recognition accuracy [i.131].

5.5.14.3 Database 34

This real-life corpora SAFE Corpus illustrates everyday life contexts in which social emotions frequently occur. The type of emotional manifestations and the degree of intensity of such emotions are determined by politeness habits and cultural behaviours. The corpus focuses on the illustration of extreme fear-type emotions in rich and varied contexts. Finally, a detection system of fear emotions based on acoustic cues has been developed to carry out an evaluation. The movies make use mostly of American English (70 % of all data). In the remaining movies, actors are portraying other English (British, Irish and Canadian) or foreign accents (French, Nordic and German). This corpus provides about 400 different speakers (47 % male, 32 % female and 1 % children) [i.132].

5.5.14.4 Database 35

ExamStress is the database of Czech speech under realistic stressed conditions and presents some selected results achieved by analysing stressed speech. The database contains read and conversational speech both in neutral and stressed state of 31 male speakers. The stressed speech was recorded during final oral examinations at the Brno University of Technology. In order to quantify the stress of individual speakers the speaker's heart rate was also measured and recorded simultaneously with the speech signal [i.133].

5.5.14.5 Database 36

EMANN - Emotional database for ANN training. The 6 sentences was read by three professional actors (two female and one male). The final database contained 720 patterns (360 patterns for one-word sentences and 360 patterns for multiword sentences).Speech recording was materialized in a recording studio with professional equipment (format "wav", sampling frequency 44,1 kHz, 24 bit). The speech corpus was composed of a written text and its corresponding speech signal, both of which were used for the training of ANN. Utterances were realized for four types of emotions: anger, boredom, pleasure and sadness. MLNN and SOM were applied particularly to the utterances processing. The method is based on the idea of the musical interval. Recorded emotion speech was subjectively evaluated by four persons [i.85].

5.5.15 Summary

A score value of the success rate of emotion identification is introduced in several cases. The total average accuracy is introduced only for 7 of the databases described (out of 36 databases). The scores range between 79 % (Danish database 26) and 85 % (Spanish database 24). The English databases show scores of 79 % and 70 % for databases 7 and 8 respectively. The German database 13 shows 80 % success and 84 % success rate is described in Japanese database 20 and for Italian database 33 the success rate is 80 %. An average score value is 77,67 %.

A very interesting experimental study on vocal emotion expression and recognition and the development of a computer agent for emotion recognition is described in [i.139]. Table 12 shows the people performance confusion matrix, table 13 shows statistics for evaluators for each emotional category and table 14 shows statistics for "actors", i.e. how well subjects portrayed emotions.

Table 12: People Performance Confusion Matrix in [%]

Category	Normal	Happy	Angry	Sad	Afraid
Normal	66,3	2,5	7,0	18,2	6,0
Happy	11,9	61,4	10,1	4,1	12,5
Angry	10,6	5,2	72,2	5,6	6,3
Sad	11,8	1,0	4,7	68,3	14,3
Afraid	11,8	9,4	5,1	24,2	49,5

Table 13: Evaluators statistics in [%]

Category	Mean	Standard Deviation	Median	Minimum	Maximum
Normal	66,3	13,7	64,3	29,3	95,7
Happy	61,4	11,8	62,9	31,4	78,6
Angry	72,2	5,3	72,1	62,9	84,3
Sad	68,3	7,8	68,6	50,0	80,0
Afraid	49,5	13,3	51,4	22,1	68,6

Table 14: Actors statistics in [%]

Category	Mean	Standard Deviation	Median	Minimum	Maximum
Normal	65,1	16,4	68,5	26,1	89,1
Happy	59,8	21,1	66,3	2,2	91,3
Angry	71,7	24,5	78,2	13,0	100,0
Sad	68,1	18,4	72,6	32,6	93,5
Afraid	49,7	18,6	48,9	17,4	88,0

V.A. Petrushin compares also the success rate for neural network recognizers and for ensemble of neural network classifiers (see table 15).

Table 15: Comparison of the success rate for NN and NN-ensembles classifiers

Category	NN recognizer [%]	Ensemble of NN classifiers [%]
Normal	55 - 65	55 - 75
Happy	60 - 70	65
Angry	60 - 80	73 - 81
Sad	60 - 70	73 -83
Afraid	25 - 50	35 - 53

It is believed that these values can show the quality influence of suitable database. It is not possible to unambiguously decide about the database quality. It is necessary to judge the recordings of emotional speech which is determined for database creation. The physical properties of the acoustic wave, which are perceived as sounds, are transformed several times: first in the organ of hearing, later at the emergence of neural excitement and last in the cerebral analysis. The melody, i.e. change of a height of voice in a sentence, is very important from the point of view of a communication. Expressive changes of melody are important indicators for an emotional and voluntary attitude of a speaker. Therefore, the sounds perceived in listening tests do not correlate completely with the objective properties of the acoustic patterns. The listening tests shall consequently form a part of experiments.

It is mandatory to apply results from the described experiments with emotional speech to the improvement of synthetic speech naturalness, but also to the domain of neurodevelopmental disturbances (above all, developmental dysphasia).

The complete overview of available speech corpora related to emotion detection including their technical details and methods employed is contained in archive ts_103296v010101p0.zip which accompanies the present document.

5.6 Spoken Emotional Speech Pre-processing

5.6.1 Preface

The speech features that efficiently characterize the emotional content of speech are categorized in four classes. They are continuous speech features, qualitative speech features, spectral features and nonlinear TEO-Based features. They are components of the speech pre-processing.

5.6.2 Introduction

Various researchers confuse terms features and parameters. The fundamental frequency F_0, intensity and duration of speech units are basic parameters for emotion classification or emotion recognition. Also voice colour (timbre) and articulation mode are important for emotional speech.

5.6.3 Features

Typical features are the pitch (fundamental frequency of the phonation F_0, pitch frequency), the formants (many algorithms for estimating the pitch signal and formants exist [i.133]), the vocal tract cross-section areas, the mel-frequency cepstral coefficients (MFCC), the Teager energy operator-based features, the intensity of the speech signal, the speech rate and the number of harmonics. Vocal tract features are very important. The shape of the vocal tract is modified by the emotional states. Many features have been used to describe the shape of the vocal tract during emotional speech production. Such features include the formants which are a representation of the vocal tract resonances, the cross-section areas when the vocal tract is modelled as a series of concatenated lossless tubes [i.136] and the coefficients derived from frequency transformations. The other features can be based on a parametrization method. They are MFCC or LFPCs (log-frequency power coefficients). Better features than MFCCs for emotion classification are in practice the LFPCs which include the pitch information [i.135]. The LFPCs are simply derived by filtering each short-time spectrum with 12 bandpass filters having bandwidths and centre frequencies corresponding to the critical bands of the human ear [i.136]. The short-term speech energy can be exploited for emotion recognition, because it is related to the arousal level of emotions. The motional force, paralinguistic effect, intonation, word juncture, the verbal content and miscellaneous noises are use in [i.137].

5.6.4 Parameters

Parameters for emotional speech processing are various. V. A. Petrushin [i.138] employ the following acoustical variables: fundamental frequency F_0 (also other authors used this one), energy, speaking rate, first three formants (F1, F2 and F3) and their bandwidths (BW1, BW2 and BW3) and calculated some descriptive statistics regarding them. Mean, median, standard deviation, maximum, minimum, range "max-min", linear regression coefficient of F_0, duration (speech-rate, ratio of duration of voiced and unvoiced region, duration of the longest voiced speech), formants (F1, F2, their bandwidths BW1, BW2 - mean of each features) use [i.141]. The speaking rate was calculated as the inverse of the average length of the voiced part of utterance. For all other parameters the following statistics were calculated: mean, standard deviation, minimum, maximum and range. Additionally for F_0 the slope was calculated as a linear regression for the voiced part of speech, i.e. the line that fits the pitch contour. The relative voiced energy as the proportion of voiced energy to the total energy of utterance was used too. Altogether they have estimated 43 features for each utterance. The other authors worked with voice intensity, cepstral coefficients, mel-frequency cepstral coefficients, F_0 envelope, Zero-Crossing Rate, pauses, duration of utterances. Table 16 summarizes the most general correlates to emotions in speech.

Table 16: Summary of most general correlates to emotions in speech. Adopted from [i.146].

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch average	Very much higher	Much higher	Slightly slower	Very much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice quality	Breathy, chest tone	Breathy, blaring	Resonant	Irregular voicing	Grumbled chest tone
Pitch changes	Abrupt, on stressed syllables	Smooth, upward inflections	Downward inflections	Normal	Wide downward terminal inflections
Articulations	Tense	Normal	Slurring	Precise	Normal

It is possible to combine several viewpoints for review of the quality of emotional databases. They are the quantitative values of parameters by a signal processing theory, but also subjective evaluation by listeners (see table 16 for example). Measurable parameter values are not introduced in [i.146].

5.6.5 Methods and materials

Methods which are frequently used in emotion recognition are described hereafter. Mainly they are standard DSP methods. PRAAT is a free scientific computer software package for the analysis of speech in phonetics. The other classification techniques can be divided into several categories, namely those employing prosody contours, i.e. sequences of short-time prosody features, statistics of prosody contours, transcription, like the mean, the variance, or the contour trends. Also three emotion classification techniques based on training of features were found in the literature, namely a technique based on artificial neural networks (ANNs) [i.140], the multichannel hidden Markov Model [i.141] and the mixture of hidden Markov models [i.142]. Classification techniques are based on artificial neural networks [i.107], Fisher linear discriminant analysis [i.143], k-nearest neighbours [i.144], kernel regression [i.144], maximum likelihood Bayes classification [i.144], linear discriminant classifiers (LDC) with Gaussian class-conditional [i.139], support vector machines [i.100]. Biomedical techniques, such as laryngograph, EEG, myogram of face, galvanic resistance and heart beat rate are applied for emotions processing.

The output of emotion classification techniques is a prediction value (label) about the emotional state of an utterance.

The different materials for emotion analysis, recognition and synthesis are used. They are (for example):

- words, numbers, sentences (affirmative, interrogative), short utterances, 10 seconds to 60 seconds clips from TV, real interviews, telephone messages (15 seconds to 90 seconds);
- computer commands, mathematics, full digit sequences, monologue;
- months, dates, special words (brake, left, slow, fast, right) - in army;
- sentence - artificial language, nouns, pseudo words, emotion intensities per emotion;
- the subjects were told to recall an emotional situation of their life and speak about that;
- unobtrusive videotaping of passengers at lost luggage counter followed up by interviews.

6 Requirements for Emotion Detectors used for Telecommunication Measurement Applications and Systems

6.1 General considerations

6.1.1 Introduction

As already described in clause 4.5 of the present document, there are two different classifications that are used to classify emotions in general, namely categorical classification and dimensional classification. The latter is very complex, implying a lower reliability of emotion detection than the categorical classification but it can be still useful for the purpose of emotion detection in the context of telecommunication measurement applications and systems as it provides valuable information for the prospective user.

On the other hand, both the categorical and dimensional classification are still a bit more complex than required as for telecommunication measurement purposes only information whether the sample of interest can evoke any emotion over subjective testing is needed. Such emotions can finally affect an assessment done by the test subjects in a subjective test and by that also bias the final subjective scores coming from the corresponding test. Such a bias could impact important decisions about a system or service design, a system or service optimization or even the training of an objective model to monitor a quality provided by that system or service in a real-life situation.

This is to say that emotion detectors trained by subjective assessments obtained from experiments based on the categorical and dimensional categorization provide the prospective user with complex information about all the emotions evoked by the corresponding text or speech sample. It is worth noting here that the corresponding information will very likely to some extent decrease the reliability of the emotion detection process as this approach is more complex, i.e. harder to execute in a reliable way (not easy to clearly classify emotions into the defined categories or correctly assign the emotion dimensions as some of them may overlap to some extent), than a simple approach indicating if the sample carries any emotion at all. Moreover, when it comes to the dimensional categorization, reliability is further decreased due to its more complex nature. The simple approach provides, in fact, all the important information when it comes to a sample selection or verification in the context of telecommunication measurement applications and systems. On the other hand, the additional information provided by these kind of detectors can be definitely very useful at the later stage of the selection or verification process, since it can be considered as diagnostic information and can be used by the prospective user to fine tune the sample to become neutral in the context of emotions and make it suitable for subjective and objective testing in the context of telecommunication measurement applications and systems.

To sum up, a preferred approach here is to use the detector deploying the simple emotion detection approach implying a higher reliability of emotion detection than the approach based on the categorization or dimensional classification. Anyhow, if such a detector is not available to the prospective user, the user is recommended to deploy a detector based on the categorization or dimensional classification, bearing in mind that this will likely affect to some extent a reliability of the emotion detection in a negative way. On the other hand, this approach offers the prospective user some diagnostic information as an added value. It is worth noting here that the extent of the negative impact on the reliability of emotion detection highly depends on the number of emotions or emotion dimensions to be classified by the detector, i.e. more emotion categories or dimensions covered by the detector imply a greater extent of the reliability decrease.

As mentioned in clause 4.6 of the present document, there are three different methods to be deployed in computer-based emotion detection, i.e. keyword-based/lexical-based, learning based and hybrid method. The first of them, namely keyword based/lexical-based method, is not convenient for detecting emotions in samples to be used in telecommunication measurement applications and systems due to two different reasons:

- First of them is that this approach is limited to a small group of the keywords covered by the databases used for the training of the detector. It should be noted here that when the group is big enough for an application scenario of interest of the prospective user, what is mostly not a case, detectors deploying this approach can become convenient for the purpose of emotion detection in the context of telecommunication measurement applications and systems, if the second reason specified below can be solved by other means.

- The second reason is that this approach is not able to recognize emotions containing no obvious affect words. In the other words, this kind of detectors is not able to detect a context of the message carried by the corresponding text or speech and therefore it is not, for instance, able to recognize sarcasm involved in the message.

Consequently, only a learning based and hybrid approach are of interest here, if the databases used in the context of keyword-based/lexical- based approach covers a reasonable amount of the keywords for the corresponding application scenario and the context of the message carried by the corresponding text or speech is properly identified by other means. It is worth noting here that the learning based approach deploys machine learning techniques, like SVM, NBC and HMM. These techniques represent the most popular ones currently deployed in automatic emotion detection. The interested reader is referred to clause 4.6.3 to get more information about their drawbacks and advantages.

Three different options to classify emotions are recommended above in the context of telecommunication measurement applications and systems:

- simple emotion detection providing a simple answer if any emotion at all is carried by the corresponding text or speech sample;

- emotion detection based on the categorization classification;

- emotion detection based on the dimensional classification.

The first one implies a very simple one-dimensional output vector, so better to say a scalar, containing binary information about presence or non-presence of any emotion at all in the corresponding text or speech sample. This information can be very easily processed automatically by a telecommunication measurement system or application. The second and third one, the categorization and dimensional classification based emotion detection, are supposed to provide the prospective user with additional data, to be prospectively treated as the diagnostic information. The output vector is more complex in this case. Such a vector is a multidimensional binary or real value vector, where the number of emotion categories or dimensions covered by the detector fully corresponds to its degree of multidimensionality. Therefore, the automatic processing of this information by a telecommunication measurement application or system requires more effort than in the previous case, of course at the avail of getting more complex data. The application or system can even provide the prospective user with some hints how to fine tune the sample/samples to fulfil all the defined criteria.

Finally, a most important parameter to consider, while selecting an appropriate detector for the purpose of emotion detection in the context of telecommunication measurement applications and systems, is naturally a measure characterizing the reliability of the information provided by the emotion detector. As already described above, there are three different approaches when it comes to the emotion detectors in the context of telecommunication measurement applications and systems, i.e. the simple emotion detection approach providing the prospective user with a binary information about presence or non-presence of any emotion at all in the text or speech sample and the approach based on the categorization or dimensional classification method providing the prospective user with a detailed multidimensional information about a group of emotions or dimensions considered in the particular categorization or dimensional approach. As the above-mentioned approaches differ quite substantially, at least the simple one and the other two ones, measure or measures to be prospectively used to characterize the reliability of the information provided by the emotion detector also vary quite substantially over these two different approaches.

When it comes to the simple method, only one measure is typically considered, i.e. success rate. The success rate is in fact just a fraction or percentage of success among some attempts. So, in the other words, it tells the prospective user how successful the particular detector is in its task.

On the other hand, the following measures are mostly used to characterize the reliability of the information provided by an emotion detector for the approaches based on the categorization or dimensional classification:

- Success rate.
- F-measure.
- Pearson correlation coefficient.

The first two measures are used in the cases when the detector provides only a binary information about a presence or non-presence of the emotions or dimensions involved in the particular categorization or dimensional approach and not information about the strength of each of the emotions or dimensions covered. Both measures are reported for each of the emotions or dimensions involved in the categorization or dimensional classification separately. On the other hand, the third one is deployed when the detector provides more complex information about the emotions or dimensions, not only binary information about a presence of non-presence of the emotions or dimensions in the text or speech sample, but also for instance strength of each of the emotions or dimensions of emotions carried by the particular text or speech sample and expressed on a similar scale as the MOS scale deployed in speech and video quality assessments. Moreover, the correlation coefficient is separately reported for each of the emotions or dimensions covered by the particular categorization or dimensional classification.

The F-measure is a measure widely known and deployed in pattern recognition and any classification tasks deploying a binary classification approach. In principle, this measure is based on two different sub-measures, i.e. precision and recall. The precision is a fraction of retrieved instances that are relevant, while recall is a fraction of relevant instances that are retrieved. In the other words, the precision provides the prospective user with an answer to the question "How many selected/identified items/features are relevant?" and the recall answers the following question "How many relevant items/features are selected/identified?". It is worth noting here that both the precision and recall are therefore based on a relevance measure [i.149]. So, the recall is of higher relevance when it comes to emotion detection in the context of telecommunication measurement applications and systems. When the above-mentioned information is put into the context of classification tasks to make it clearer for the prospective reader of the present document, the categorization and dimensional classifications are fully based on this principle, an output provided by the detector under test is compared with trusted external judgments by deploying the following terms true positive (*tp*), true negative (*tn*), false positive (*fp*) and false negative (*fn*), see Type I and type II errors for the corresponding definitions [i.150].

The terms positive and negative refer to a detector's prediction (sometimes known as an expectation) and the terms true and false refer to whether that prediction corresponds to the external judgment (sometimes known as an observation). As a result of this process, the precision and recall can be defined as follows:

$$Precision = \frac{tp}{tp+fp} \times 100\,\% \tag{6.1}$$

$$Recall = \frac{tp}{tp+fn} \times 100\,\% \tag{6.2}$$

The recall in this context is also referred to as a true positive rate or sensitivity and the precision is also referred to as a positive predictive value. Another related and relevant measure used quite frequently in classification tasks is accuracy, which is defined as follows:

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \times 100\,\% \tag{6.3}$$

A measure that combines the precision and recall is called F-measure, represents a harmonic mean of the precision and recall and is defined as follows:

$$F = 2 \times \frac{precision \times recall}{precison+recall}\ \% \tag{6.4}$$

The Pearson correlation coefficient is well known in the speech and video quality assessment community and is defined as follows:

$$R = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}} \times\ 100\,\% \tag{6.5}$$

NOTE: For the purposes of the present document, R is expressed in percent.

where X_i is a value of the parameter characterizing for instance a strength (or presence) of emotion carried by the text or speech sample or other parameter quantifying an amount of the particular emotion carried by the text or speech sample obtained from a subjective test for corresponding configuration/scenario *i*, Y_i is a value of the same parameter estimated by the emotion detector, $\overline{X}$ and $\overline{Y}$ represent the corresponding arithmetic mean values, *N* is a number of stimuli considered in the comparison. It can be applied both in the case of simple detectors providing binary output (emotion present/not present) and in the case of continuous-scale output. This measure is frequently accompanied by root mean square error and other similar measures to be frequently deployed also in speech or video quality assessment see Recommendation ITU-T P.1401 [2] for more detail, to provide the prospective user with more information about the performance of the detector.

6.1.2 Computational Power Requirements

It needs to be mentioned here that a testing phase, which is of most interest here, does not require a large amount of computational resources, e.g. computation time, memory space, etc., as a training phase of the detector. Therefore, the type of machine learning technique deployed by the detector should not also be of high importance when selecting an appropriate detector for the purpose of telecommunication measurement applications and systems. So, this means that a machine to be used for running an emotion detector in the testing phase does not require a huge amount of the computational resources. Hence, a standard consumer computer regarding CPU and RAM should be convenient for this purpose. On the other hand, a very powerful computer is required to run the training phase of emotion detector. The prospective user should also keep in mind in this case that the final performance of the detector is not uniquely influenced by the machine learning technique deployed, but also by the databases used in the training process. Therefore, a reasonable amount of information about the databases involved in the training process of the detector should also be provided by the vendor of the emotion detector to allow the prospective user to properly identify the best possible detector to fulfil all the expectations and cover all the application scenarios defined by the prospective user.

The algorithm performing in accordance with the present document shall specify the memory and computational power requirements for the operational phase. Corresponding requirements for the training phase should be provided too.

6.1.3 Requirements for Operational Modes

Emotion detectors can be classified into two categories when it comes to types of operation. The first category contains the following operation modes:

- real-time;
- quasi-real-time (noticeable delay caused by the process, which is still acceptable for the rest of the process or other devices involved);
- off-line.

The second category covers the following types of operation:

- standalone (a device using dedicated hardware and software);
- embedded (installed as a part of other device/application which cannot work separately);
- modular application (installed on an independent platform and acting as a separate device).

When it comes to minimum requirements, in this case, the prospective user of emotion detector should be able, on the basis of the information provided by the vendor of that device/algorithm, to assign that device/algorithm to one of the operation modes or types of operation of both categories listed above. So, the information provided by the vendor should contain all the information needed to identify clearly, or should clearly indicate, types of operation of emotion detector in both cases.

6.2 Requirements for Emotion Detectors for Written Text

First of all, it shall be properly checked whether the application scenario of an emotion detector for written text, which is going to be deployed in the context of emotion detection for telecommunication measurement applications and systems, matches the application scenario required by the prospective user.

It shall be confirmed that the context wherein the emotion detector is supposed to indicate if any emotion at all is carried by the corresponding text (the simple approach) or even identify and categorize emotions or emotion features carried by the corresponding text into the particular emotion categories or dimensions covered by the detector (the approaches based on the categorization and dimensional classification) is fully covered by the scope of that specific emotion detector.

It shall be confirmed that the detector is properly trained to be able to recognize emotions in the corresponding application scenario and context of use. Databases used in the training and verification process of that specific emotion detector shall fully cover the particular application scenario and context of use.

Moreover, types of operation of the detector should also be clearly indicated by the information provided by the vendor. All the above-mentioned requirements represent processing requirements of emotion detector for written text.

When it comes to input requirements of emotion detector for written text, a text, which is supposed to undergo emotion detection, should be provided in a sentence form, as this form is of interest, when it comes to telecommunication measurement applications and systems and a text format required by the corresponding emotion detector.

As in the other related fields, e.g. speech and video quality assessment, a quality estimation or prediction model is considered as reliable if a value estimated/predicted by the model correlates with a ground truth or in other words with a quality value obtained from the subjective test on a level above 90 % (provisional value) expressed by the Pearson correlation coefficient. As the output of emotion detectors is supposed to be used in a selection or verification process of test material going to be deployed in subjective or objective quality testing, a reliability level outlined above for speech and video quality assessment shall also be met for emotion detectors for written text in order to keep a reasonable level of reliability in all the processes involved in the preparation and execution part of a quality assessment process. Nevertheless, lower values of the reliability level can still improve results coming from a subjective testing regarding the negative impact of emotions carried by the test material on subjective quality scores.

An emotion detector for written text shall have a reliability, representing an output requirement of emotion detector for written text, higher than 90 % (provisional value), expressed by any of the above listed measures, e.g. success rate, F-measure (precision, recall) or Pearson correlation coefficient in the context of all the approaches to be prospectively used for emotion detection for telecommunication measurement applications and systems, namely the simple one and approaches based on the categorization and dimensional classification.

Table 17 summarizes the measures to be prospectively used to characterize the reliability of emotion detector for written text in the context of telecommunication measurement applications and systems and all the prospective approaches together with the required reliability level.

Table 17: Summary of reliability measures and requirements for emotion detectors for text

Approach	Measure	Minimum value (provisional value)
Simple	Success rate	> 90 %
	Pearson correlation coefficient	> 90 %
Categorized	Success rate	> 90 %
	F-measure (precision, recall)	> 90 %
	Pearson correlation coefficient	> 90 %
Dimensional	Success rate	> 90 %
	F-measure (precision, recall)	> 90 %
	Pearson correlation coefficient	> 90 %

Table 18 presents a summary of all the requirements defined for emotion detectors for text together with their status.

Table 18: Summary of all the requirements defined for emotion detectors for text

Type of requirement	Input requirements	Processing requirements				Output requirements
Name of requirement	Sentence form	Application scenario	Context of use	Proper training	Types of operation	Parameters as per table 17
Status	Mandatory	Mandatory	Optional	Optional	Optional	Mandatory

6.3 Requirements for Emotion detectors for speech

Firstly, it shall be properly checked whether the application scenario of an emotion detector for speech, which is going to be deployed in the context of emotion detection for telecommunication measurement applications and systems, matches the application scenario required by the prospective user.

It shall be confirmed that the context wherein the emotion detector is supposed to indicate if any emotion at all is carried by the corresponding speech signal (the simple approach) or even identify and categorize emotions or emotion features carried by the corresponding speech signal into the particular emotion categories or dimensions covered by the detector (the approaches based on the categorization and dimensional classification), is fully covered by the scope of that specific emotion detector.

It shall be confirmed that the detector is properly trained to be able to recognize emotions in the corresponding application scenario and context of use. Databases used in the training and verification process of that specific emotion detector shall fully cover the particular application scenario and context of use. Since the detection of emotions carried by speech is a much more complex task than that of dealing with text only, it should be taken into account, while selecting an appropriate emotion detector, that the creation of an emotion database constitutes a very important role here. In the other words, there are, as also clearly described in clause 5.2, three completely different approaches how to create an emotion database for the training of an emotion detector:

- using professional actors - acted emotions;
- using a naïve subject and external stimulus - provoked emotions;
- using spontaneous emotions recorded in real-life situations.

It should be noted here that the last approach provides most reliable results as it fully reflects real emotions. On the other hand, this type of databases is very difficult to create. So, the prospective user of detector of emotions carried by speech is highly advised to focus only on detectors being based on databases involving the spontaneous emotions recorded in a real-life situations (a preferred option) or databases containing the provoked emotions (a kind of back-up option) in order to get realistic and reliable output data in the context of telecommunication measurement applications and systems.

Moreover, types of operation of the detector should also be clearly identified or even clearly indicated by the information provided by the vendor. All the above-mentioned requirements represent processing requirements of emotion detector for speech.

Concerning input requirements of emotion detector for speech, speech supposed to undergo emotion detection should be provided in a sentence form, as this form is of interest in the context of telecommunication measurement applications and systems, and a speech format required by the corresponding emotion detector. Moreover, speech signals deployed for emotion detection process should be full band audio signals, having preferably a sampling rate of 48 kHz and bit resolution at least of 16 bit per sample in order not to bias the emotion detection process by distortions introduced by bandwidth limitation. As in the other related fields, e.g. speech and video quality assessment, a quality estimation or prediction model is considered as reliable if a value estimated/predicted by the model correlates with a ground truth or in other terms with a quality value obtained from the subjective test on a level above 90 % (provisional value), expressed by the Pearson correlation coefficient. As the output of emotion detectors is supposed to be used in a selection or verification process of test material going to be deployed in subjective or objective quality testing, a reliability level outlined above for speech and video quality assessment shall also be met for emotion detectors for speech in order to keep a reasonable level of reliability in all the processes involved in the preparation and execution part of a quality assessment process. Nevertheless, lower values of the reliability level can still improve results coming from a subjective testing in terms of the negative impact of emotions carried by the test material on subjective quality scores.

An emotion detector for speech shall have a reliability, representing an output requirement of emotion detector for speech, higher than 90 % (provisional value), expressed by any of the above listed measures, e.g. success rate, F-measure (precision, recall) or Pearson correlation coefficient in the context of all the approaches to be prospectively used for emotion detection for telecommunication measurement applications and systems, namely the simple one and approaches based on the categorization and dimensional classification.

Table 19 summarizes the measures to be prospectively used to characterize the reliability of emotion detector for speech in the context of telecommunication measurement applications and systems and all the prospective approaches together with the provisional reliability level.

Table 19: Summary of reliability measures and requirements for emotion detectors for speech

Approach	Measure	Minimum requirement (provisional value)
Simple	Success rate	> 90 %
Categorized	Success rate	> 90 %
	F-measure (precision, recall)	> 90 %
	Pearson correlation coefficient	> 90 %
Dimensional	Success rate	> 90 %
	F-measure (precision, recall)	> 90 %
	Pearson correlation coefficient	> 90 %

Table 20 presents a summary of all the requirements defined for emotion detectors for speech together with their status.

Table 20: Summary of all the requirements defined for emotion detectors for speech

Type of requirement	Input requirements		Processing requirements				Output requirements
Name of requirement	Sentence form	Quality of signal	Application scenario	Context of use	Proper training	Types of operation	Parameters as per table 19
Status	Mandatory	Mandatory	Mandatory	Optional	Optional	Optional	Mandatory

6.4 A combined method and its requirements

6.4.1 General description

This method represents a two-stage or in other terms a cascade approach. At the first stage, emotions carried (if any) by the text are supposed to be detected by an emotion detector for text. This part shall remove a reasonable portion of emotions carried by the text. So, this may also rapidly reduce the time spent or required for recording test speech signals to be used by telecommunication measurement applications and systems as the selected text material shall be mostly free of any emotions. Anyhow, this naturally depends on the reliability level of the detector deployed for this purpose. In the second stage, the recorded speech material needs to be checked in order to verify if it carries any emotions by an emotion detector for speech. This part is supposed to eliminate as much as possible the speech material carrying any emotions. A result of this process also highly depends on the reliability level of the detector used in this case.

6.4.2 Requirements of the combined method

The cascade approach deployed in this method softens the output requirement for both the emotion detectors, namely detector for text and detector for speech, being a part of this method. However the reliability level of the overall emotion detection process defined already above separately for the emotion detectors for speech and text remains the same. So, the corresponding output requirement can be achieved at lower costs in terms of reliability of both the involved emotion detectors. Table 21 summarizes all the measures to be prospectively used to characterize the reliability of the combined method in the context of telecommunication measurement applications and systems and prospective approaches and their combinations together with the provisional reliability level of the overall emotion detection process. It should be noted here that all the other requirements, i.e. the input and processing requirements listed above for text and speech emotion detectors separately, apply appropriately also for the combined method in general or its parts.

Table 21: Summary of reliability measures and requirements for the combined method

Approach (Text/Speech)	Measure (Text/Speech)	Minimum requirement (provisional value)
Simple/Simple	Success rate/Success rate	> 90 %
Simple/Categorized	Success rate/Success rate, F-measure (precision, recall), Pearson correlation coefficient	> 90 %
Categorized/Simple	Success rate, F-measure (precision, recall), Pearson correlation coefficient/Success rate	> 90 %
Simple/Dimensional	Success rate/Success rate, F-measure (precision, recall), Pearson correlation coefficient	> 90 %
Dimensional/Simple	Success rate, F-measure (precision, recall), Pearson correlation coefficient/Success rate	> 90 %
Categorized/Categorized	Success rate, F-measure (precision, recall), Pearson correlation coefficient/Success rate, F-measure (precision, recall), Pearson correlation coefficient	> 90 %
Categorized/Dimensional	Success rate, F-measure (precision, recall), Pearson correlation coefficient/Success rate, F-measure (precision, recall), Pearson correlation coefficient	> 90 %
Dimensional/Categorized	Success rate, F-measure (precision, recall), Pearson correlation coefficient/Success rate, F-measure (precision, recall), Pearson correlation coefficient	> 90 %
Dimensional/Dimensional	Success rate, F-measure (precision, recall), Pearson correlation coefficient/Success rate, F-measure (precision, recall), Pearson correlation coefficient	> 90 %

Table 22 presents a summary of all the requirements defined for emotion detectors for the combined method together with their status.

Table 22: Summary of all the requirements defined for the combined method

Type of requirement	**Input requirements**		**Processing requirements**				**Output requirements**
Name of requirement	Sentence form (T, S)	Quality of signal (S)	Application scenario (T, S)	Context of use (T, S)	Proper training (T,S)	Types of operation (T, S)	Parameters as per table 21 (T, S)
Status	Mandatory	Mandatory	Mandatory	Optional	Optional	Optional	Mandatory
NOTE: T denotes emotion detector for text and S denotes emotion detector for speech.							

7 Accuracy of Emotion Detectors for Subjective Testing in Telecommunications

7.1 Introduction

This clause discusses the aspects of emotion detectors which provide binary results, i.e. the detection result is either that a text or speech sample contains emotions or that it does not contain emotions. It is seen as most likely that such binary detectors will be the first choice if it comes to the use of emotion detectors in speech sample selection. The present clause describes this particular application only.

7.2 Reference set of samples

It is important to understand that the accuracy of any emotion detector can only be assessed if the proportion of emotions contained in a reference set of text or speech samples is known with a certain interval.

In order to assess the emotional content of a reference set of samples, it is mandatory to conduct a series of subjective experiments in which a number of human subjects "detect" whether the samples have emotional content or not. There shall be different experiments designed for humans reading written text samples and for humans listening to speech samples. The design of such experiments shall follow the principles laid down in Recommendation ITU-T P.800 [4], which however has been written for speech quality assessment.

Annex B of the present document provides some high-level guidelines for the design of such experiments. However, the applicability of such experimental design for the assessment of emotions has not yet been proven. Related standards have not been made available up to now and further scientific research is required for this topic.

7.3 Assessment of the accuracy

The accuracy of an emotion detector shall be assessed by comparing the output of the emotion detector with the a-priori determined emotional content of the reference set of samples.

In a first step, false negative percentage (*fn*) and false positive percentage (*fp*) shall be determined. False negative *fn* is the percentage of emotional samples contained in the reference set of the sample that remains un-recognized by the emotion detector. False positive *fp* is the percentage of emotion-free samples in the reference set that by mistake are recognized as emotional.

False positive *fp* cannot be ignored, because it reduces the total number of samples in the reference set and thus has a direct influence on the percentage of emotional content of the sample set.

In order to achieve required assessment accuracy, the size of the sample set should be properly dimensioned, i.e. should be large enough. Common procedures for statistically minimum valid sample size should be followed [1], [2] and [3].

The accuracy of emotion detectors, both for written text and for spoken speech is described by the percentage of "Emotional Samples in Resulting Material" (*ESRM*), which is calculated according to equation (7.1):

$$ESRM = ESSM \times \frac{fn}{100 - ESSM \times (100 - fn) - (100 - ESSM) \times fp} \times 100\ \% \qquad (7.1)$$

with:

- *ESSM* = percentage of emotional samples in source material;
- *ESRM* = percentage of emotional samples in resulting material;
- *fn* = percentage of false negative detected samples;
- *fp* = percentage of false positive detected samples.

Figures 15 to 18 illustrate the dependency of *ESRM* over *ESSM* for some example combinations of *fn* and *fp*. The diagrams combine *fp* = 0 % (figure 15), *fp* = 5 % (figure 16), *fp* = 20 % (figure 17) and *fp* = 60 % (figure 18) with a variation for *fn* of 5 %, 20 %, 60 % and 100 %.

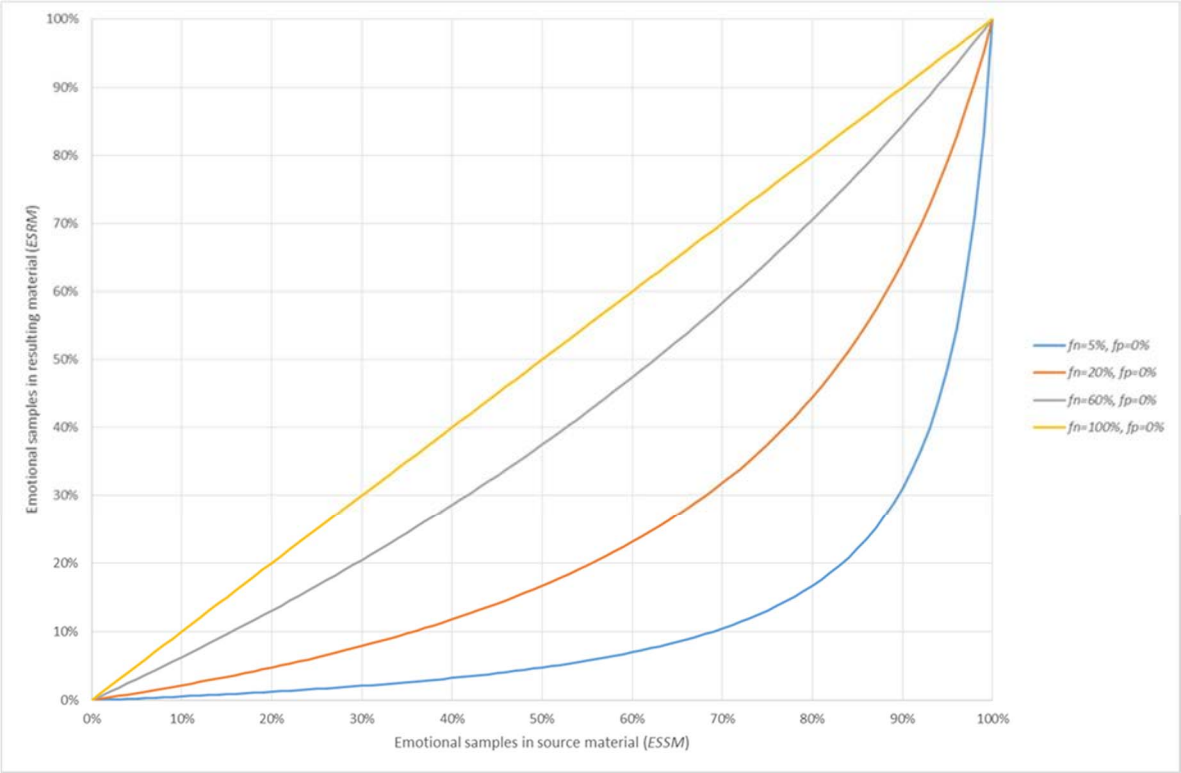


Figure 15: Example of most probable result for *ESRM* with *fp* = 0 % and *fn* as a variable parameter; comments from clause 7.3 apply

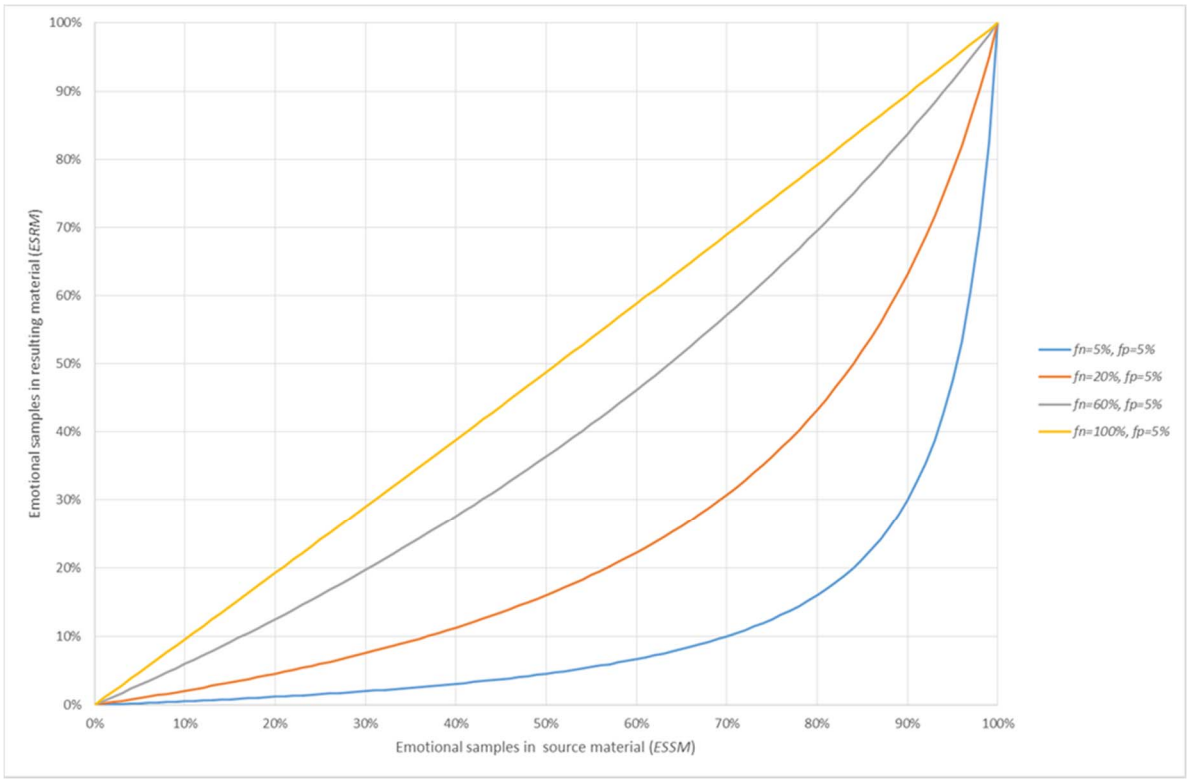


Figure 16: Example of most probable result for *ESRM* with *fp* = 5 % and *fn* as a variable parameter; comments from clause 7.3 apply

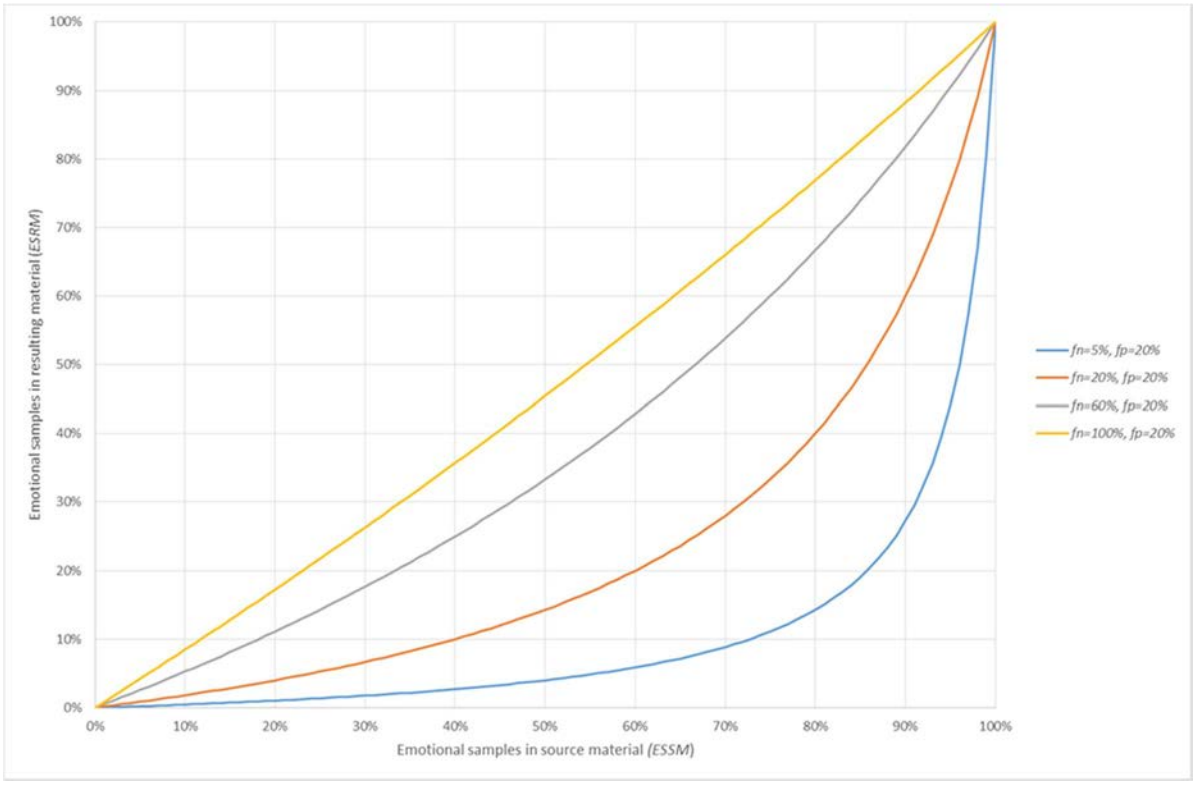


Figure 17: Example of most probable result for *ESRM* with *fp* = 20 % and *fn* as a variable parameter; comments from clause 7.3 apply

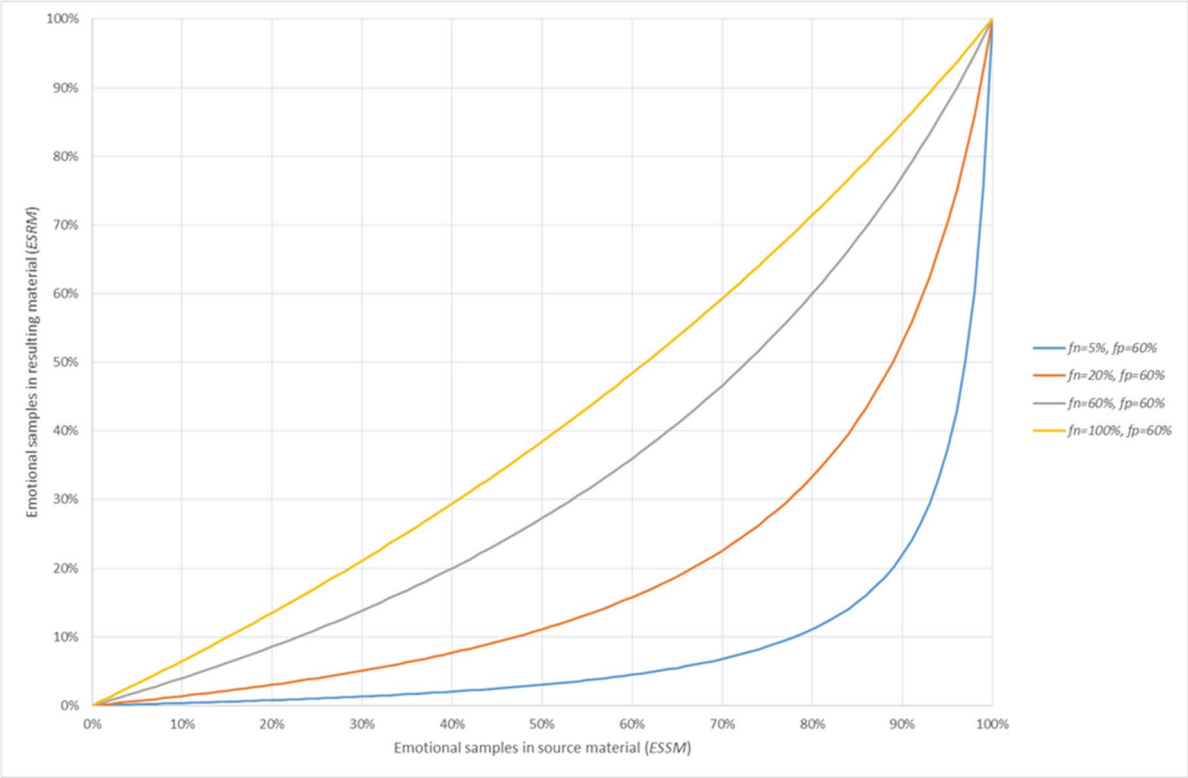


Figure 18: Example of most probable result for *ESRM* with *fp* = 60 % and *fn* as a variable parameter; comments from clause 7.3 apply

7.4 Remaining Percentage of Samples

Having in mind that one of the main purposes of the use of emotion detectors in telecommunication measurement applications is the conductance of auditive (and emotion-free) speech material, it is also very important to consider the remaining number of samples after application of the emotion detector.

This can be calculated as "Remaining Percentage of Samples" (*RPS*) according to equation (7.2):

$$RPS = 100 - \frac{ESSM \times (100 - fn)}{100} - \frac{(100 - ESSM) \times fp}{100} \ \% \qquad (7.2)$$

with:

- *RPS* = remaining percentage of samples;
- *ESSM* = percentage of emotional samples in source material;
- *fn* = percentage of false negative detected samples;
- *fp* = percentage of false positive detected samples.

Figures 19 to 22 illustrate the dependency of *RPS* over *ESSM* for some example combinations of *fn* and *fp*. The diagrams combine *fp* = 0 % (figure 19), *fp* = 5 % (figure 20), *fp* = 20 % (figure 21) and *fp* = 60 % (figure 22) with a variation for *fn* of 5 %, 20 %, 60 % and 100 %.

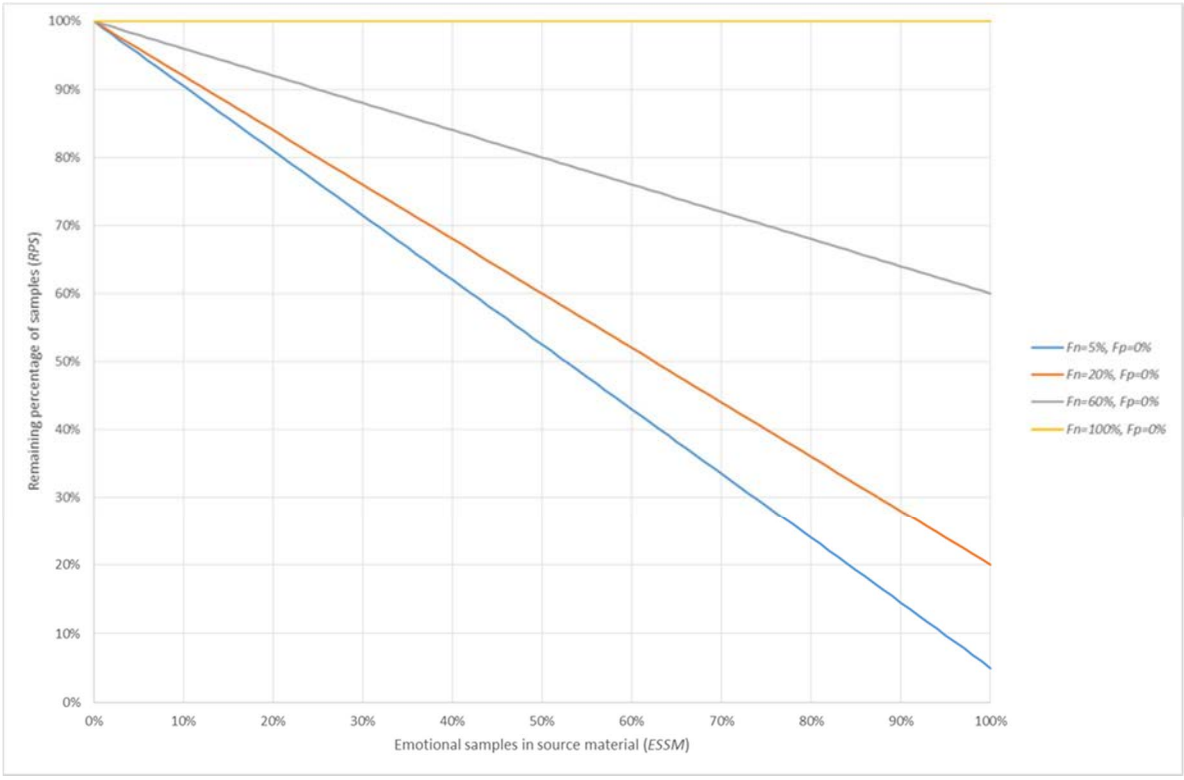


Figure 19: Example of most probable result for *RPS* with *fp* = 0 % and *fn* as a variable parameter; comments from clause 7.3 apply

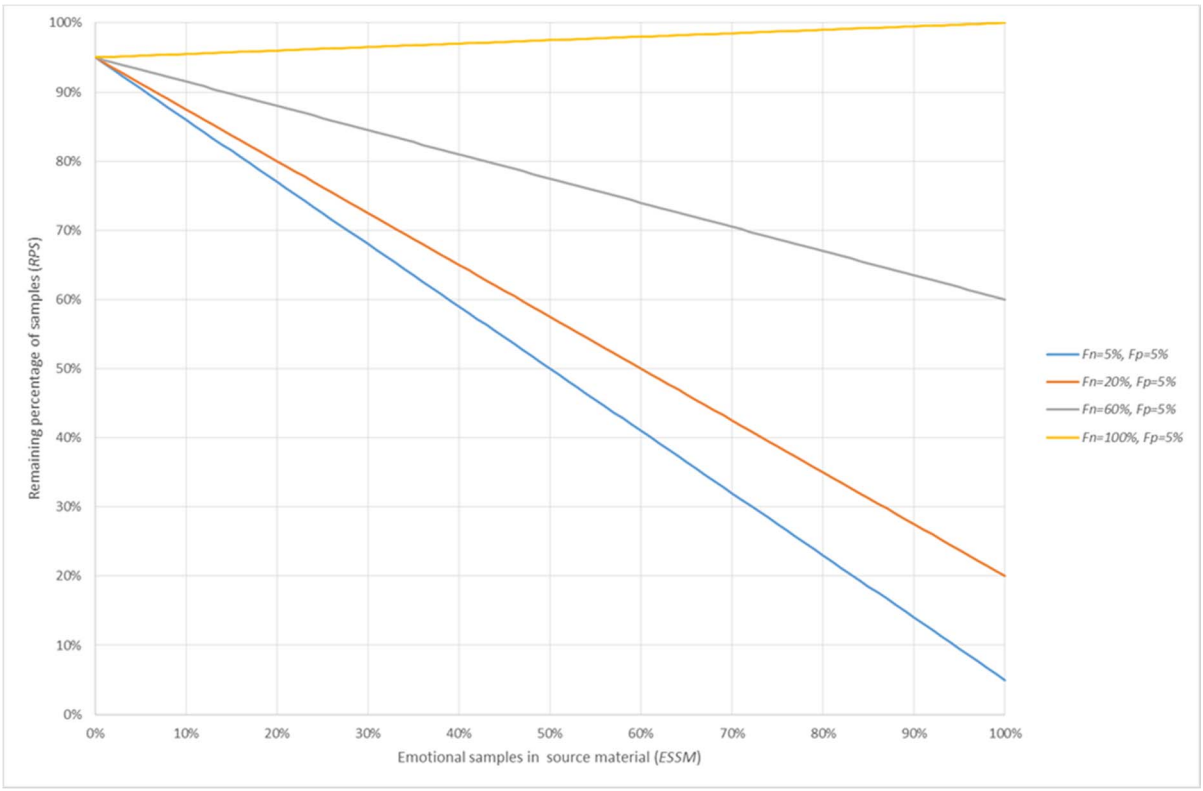


Figure 20: Example of most probable result for *RPS* with *fp* = 5 % and *fn* as a variable parameter; comments from clause 7.3 apply

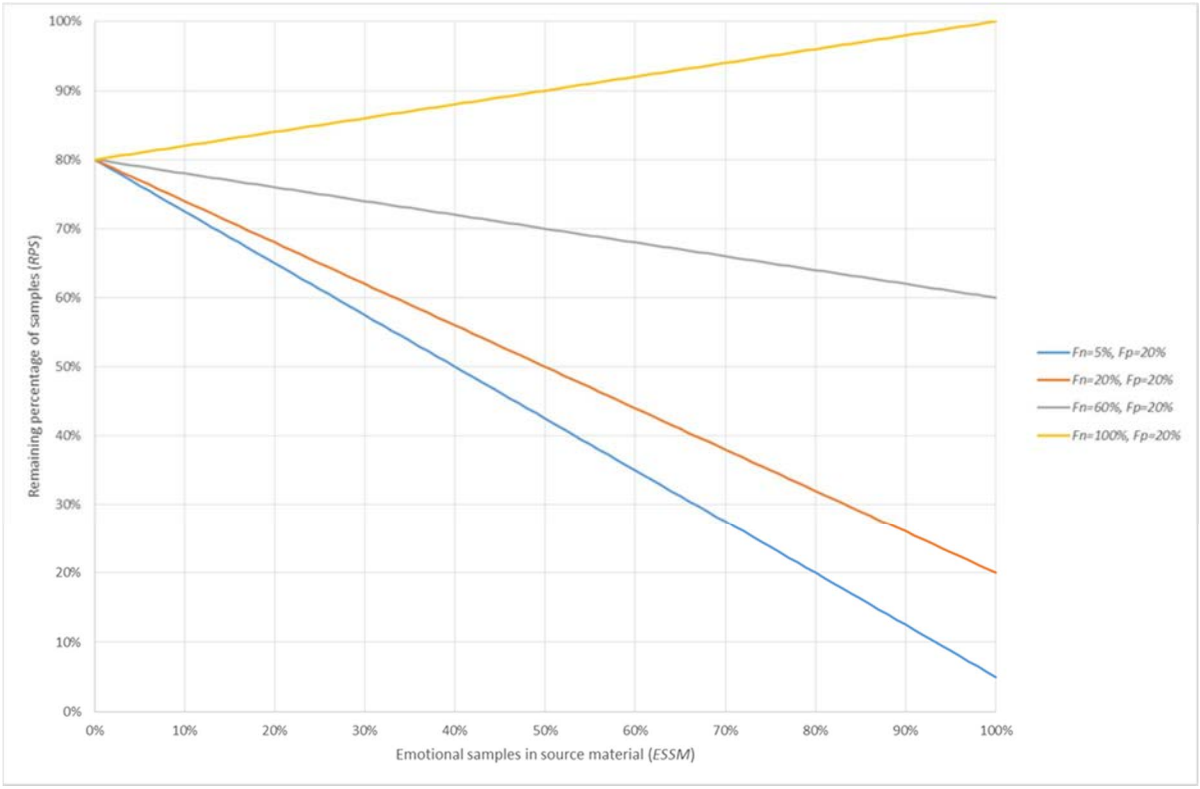


Figure 21: Example of most probable result for *RPS* with *fp* = 20 % and *fn* as a variable parameter; comments from clause 7.3 apply

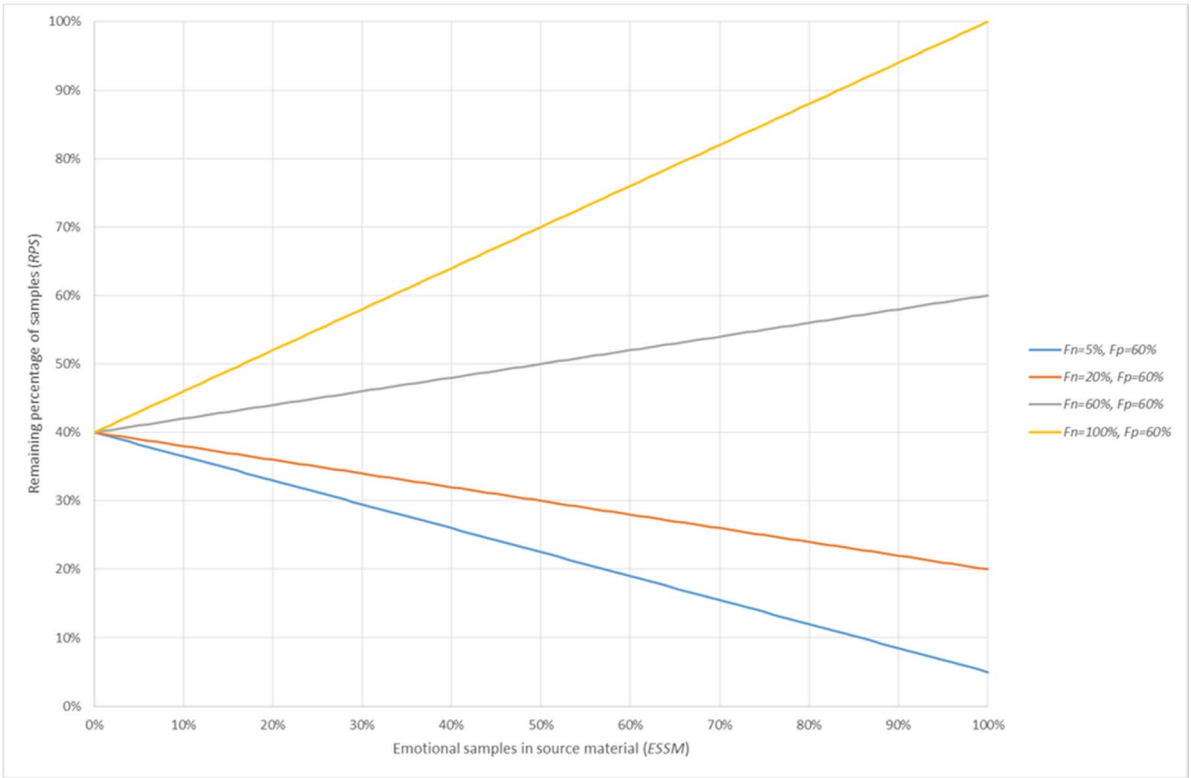


Figure 22: Example of most probable result for *RPS* with *fp* = 60 % and *fn* as a variable parameter; comments from clause 7.3 apply

7.5 Examples of Single Emotion Detectors

Figures 23 to 25 show the principles of single emotion detectors (both for written text and for spoken speech) for different combinations of *fn* and *fp*. The original sample set contains 50 samples, by coloured dots, red indicating samples containing emotions and blue indicating emotion-free samples. By applying the emotion detector, the number of samples left in the set is typically reduced. The arrows pointing upward indicate the proportion of emotions that would occur if the respective sample set is used for a subjective speech quality evaluating experiment.

The emotion detector, sitting in the middle of the figure, analyses, one-by-one the samples provided in the source sample set. Those samples, which are detected as containing emotions are removed from the sample set - depicted as being moved into the bucket.

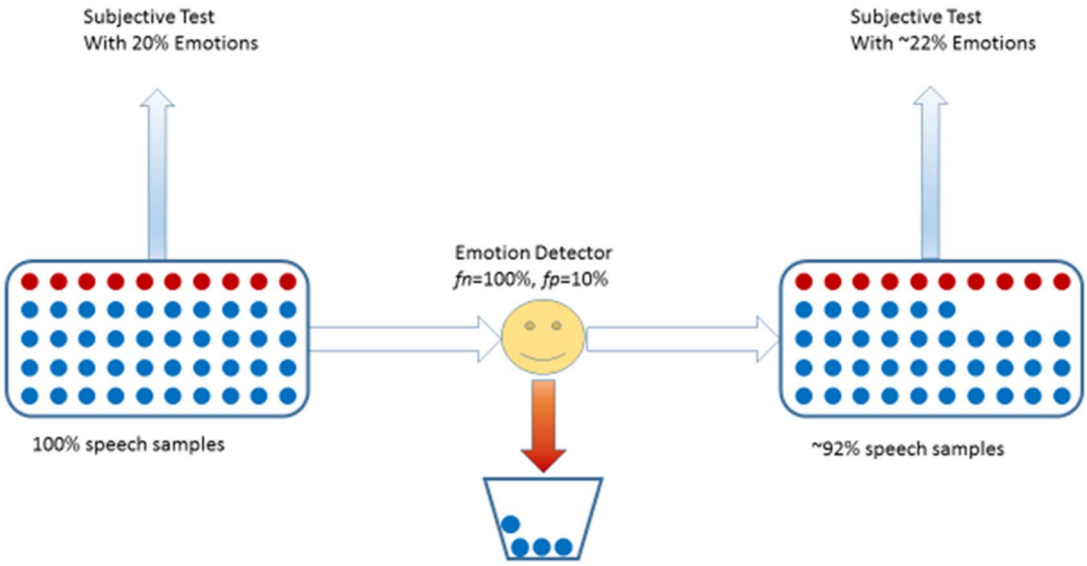


Figure 23: Example of most probable result of emotion detector with *fn* = 100 % and *fp* = 10 %; comments from clause 7.3 apply

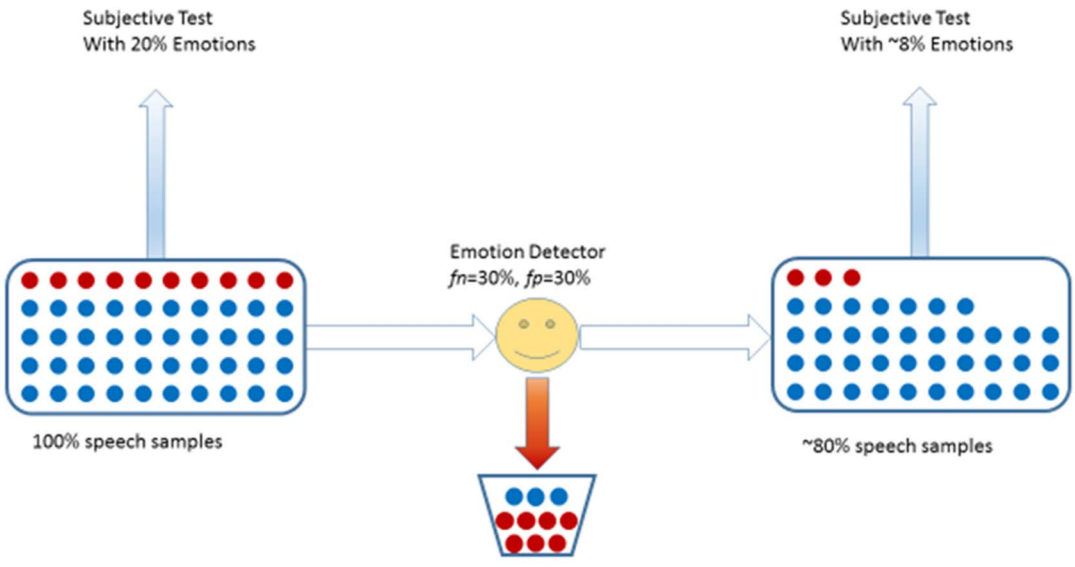


Figure 24: Example of most probable result of emotion detector with *fn* = 30 % and *fp* = 30 %; comments from clause 7.3 apply

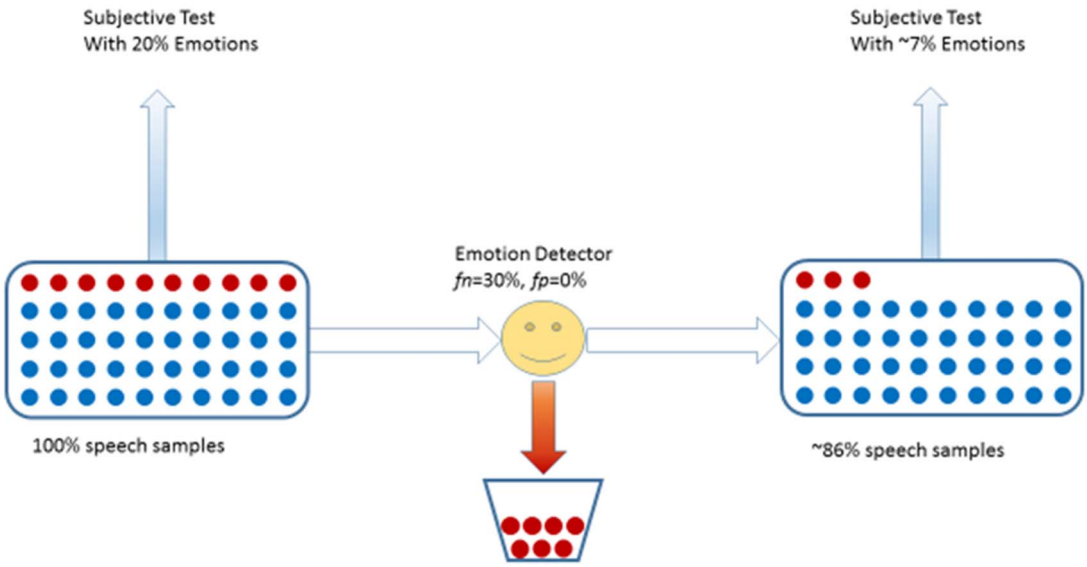


Figure 25: Example of most probable result of emotion detector with *fn* = 30 % and *fp* = 0 %; comments from clause 7.3 apply

7.6 Examples of combined emotion detectors

7.6.1 Optimum Recording process with no errors

Figures 26 to 28 show the principles of combined emotion detectors, where the first detector works for written text and the second for spoken speech for a different combination of *fn* and *fp*. As a simplification in these figures, the percentage of emotional content does not change when the set of samples is converted from written text into spoken speech.

These figures illustrate the effect of combined emotion detectors, where the recording process of the written text samples (output of the first emotion detector), takes place without any error in order to provide the input sample for the spoken speech emotion detector, i.e. the recording process does not add any emotions. This constitutes the optimum recording process.

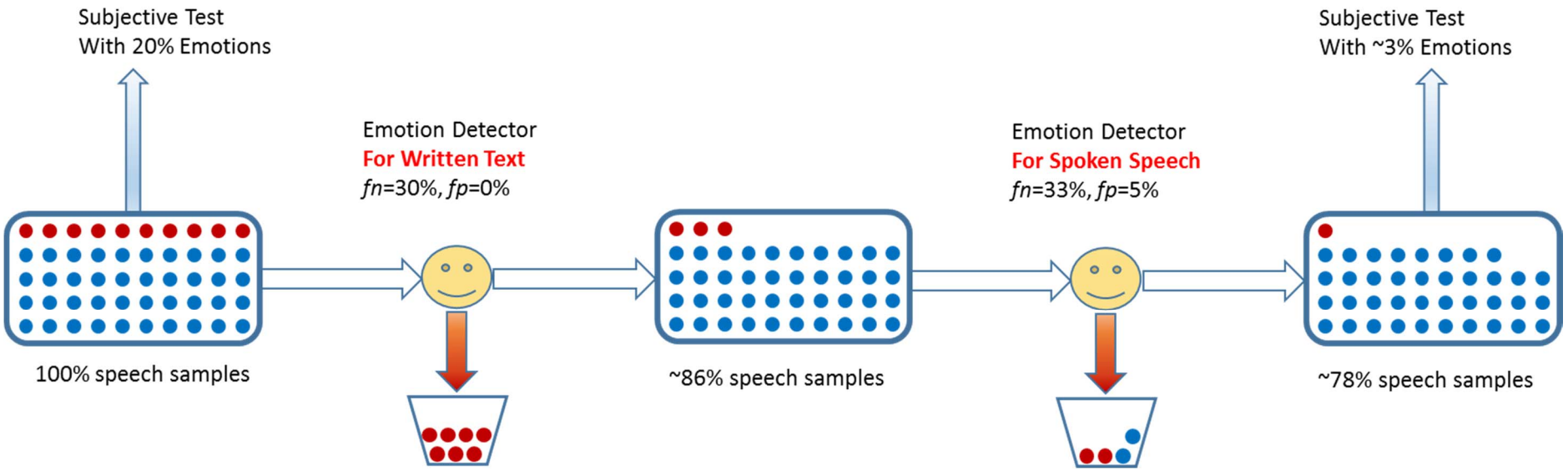


Figure 26: Example of most probable result of emotion detector for written text with *fn* = 30 % and *fp*=0 % followed by emotion detector for spoken speech with *fn* = 33 % and *fp* = 5 %; comments from clause 7.3 apply

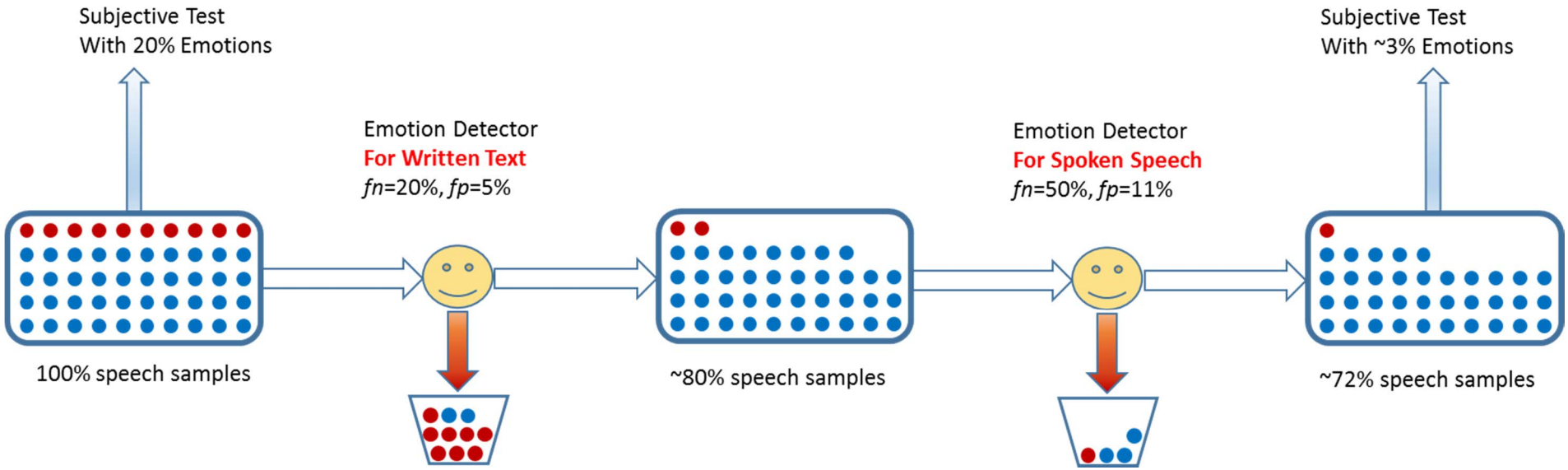


Figure 27: Example of most probable result of emotion detector for written text with *fn* = 20 % and *fp* = 5 % followed by emotion detector for spoken speech with *fn* = 50 % and *fp* = 11 %; comments from clause 7.3 apply

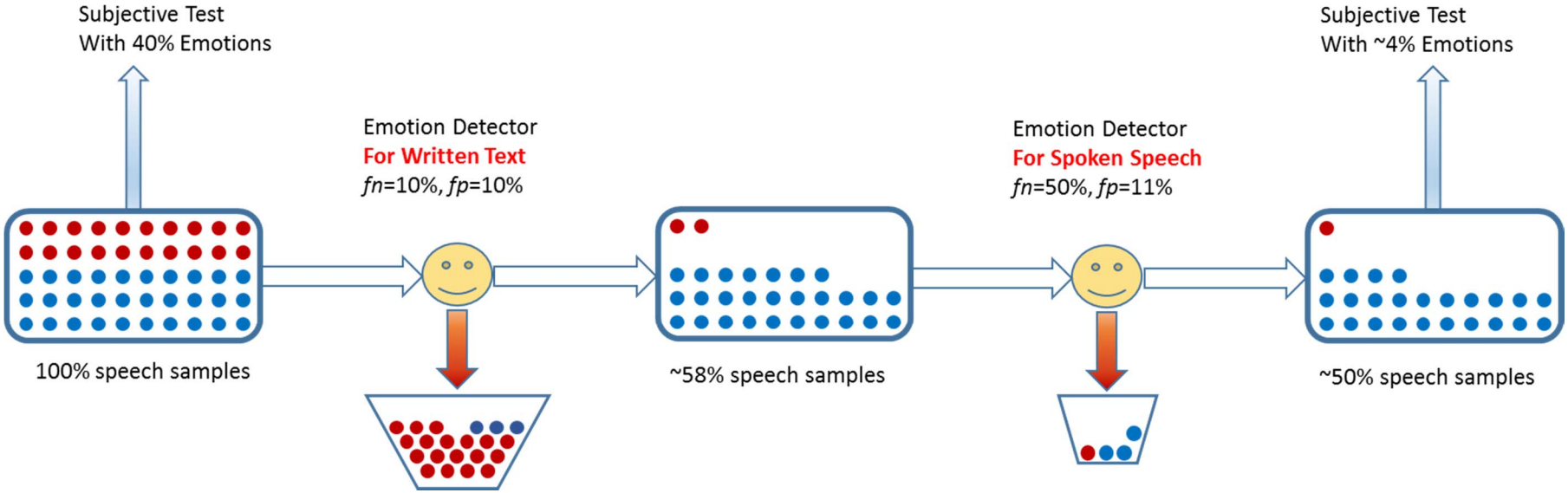


Figure 28: Example of most probable result of emotion detector for written text with *fn* = 10 % and *fp* = 10 % followed by emotion detector for spoken speech with *fn* = 50 % and *fp* = 11 %; comments from clause 7.3 apply

7.6.2 Poor Recording process with errors

Figure 29 shows the principles of combined emotion detectors, where the first detector works for written text and the second for spoken speech for a different combination of *fn* and *fp*. As indicated by the black dots another factor of uncertainty is introduced in this case when the set of samples is converted from written text into spoken speech. This may result from untrained or non-native speakers.

These figures illustrate the effect of combined emotion detectors, where the recording process of the written text samples (output of the first emotion detector), takes place with errors in order to provide the input sample for the spoken speech emotion detector, i.e. the recording process does add new emotions. This constitutes a poor recording process.

"Transition from written text into spoken speech" refers to the recording process.

"Emotion-free content converted to emotional samples by mistake" refers to the fact that during the - poor - recording process, emotion-free written text samples are converted to spoken speech samples containing emotions.

The principles stated in clause 7.6 are applicable only in case of absolute statistical independency of written and speech emotion detection procedures, which means that the application of the text emotion detection does not affect the performance of a subsequent application of a spoken emotion detector. For certain detector types, this may be considered as a quite demanding requirement and its validity has to be checked.

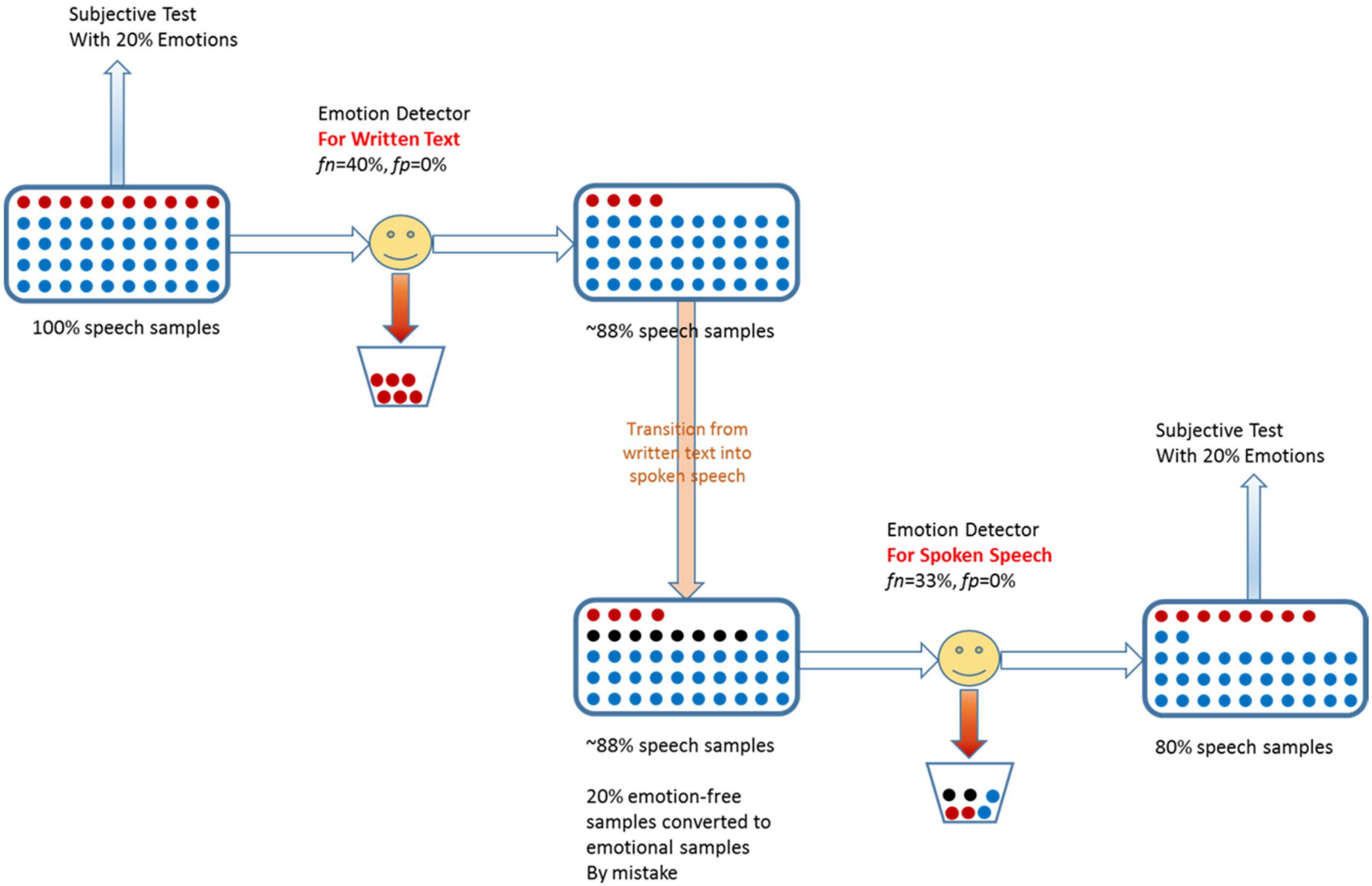


Figure 29: Example of most probable result of emotion detector for written text with *fn* = 40 % and *fp* = 0 % followed by emotion detector for spoken speech with *fn* = 33 % and *fp* = 0 %; comments from clause 7.3 apply

Annex A (informative): Overview of Available Speech Corpora

The complete overview of available speech corpora related to emotion detection including their technical details, methods employed, etc. is contained in archive ts_103296v010101p0.zip which accompanies the present document.

Annex B (informative): Subjective Assessment of Emotional Content

This annex provides guidance on aspects of subjective assessment of emotional content. It focuses on the case in which binary results are obtained, i.e. the subjects are instructed in such a way that they make their judgement whether a text or speech sample contains emotions or whether it does not contain emotions. Such binary results will be needed as a reference in the selection process of emotion detectors for telecommunication system applications.

The present clause describes this particular case, only. However, the principles remain the same for emotion detectors with scalar or multi-dimensional output and the following may be applied in a similar manner.

In order to assess the emotional content of a reference set of samples, it is mandatory to conduct a series of subjective experiments in which a number of human subjects "detect" whether the samples have emotional content or not. There will be different experiments designed for humans reading written text samples and for humans listening to spoken speech samples. The design of such experiments will follow the principles laid down in Recommendation ITU-T P.800 [i.148], which however has been written for speech quality assessment.

This annex provides some high-level guidelines for the design of such experiments. However, the applicability of such experimental design for the assessment of emotions has not yet been proven. Related standards have not been made available up to now. Therefore, additional standardization activities are required and further scientific research is advised for this topic.

The preferred test method for the subjective assessment of emotional content are "listening-only" or "read-only" tests using the "Absolute Category Rating" (ACR) method described in annex B of [i.148], which is in conformance with the Category Judgement method recommended for conversation tests (see annex A of [i.148]) and adopted partly for the same reasons.

Expert testing using a limited number of highly trained experienced subjects is also acceptable.

Annex C (informative): Bibliography

- Iida, A., Campbell, N. and Yasamura, M. (1998): "Design and Evaluation of Synthesised Speech with Emotion", Journal of Information Processing Society of Japan, 40 (2).
- Fernandez, R. and Picard, R.: "Modeling drivers' speech under stress", Speech Comm., 40 , pp .145-159.
- Murray, I. R. and Arnott, J. L. (1993): "Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature of Human Vocal Emotion", Journal of Acoustic Society of America 93 (2), pp. 1097-1198.
- Ortony, A. and Turner, T. J. (1990): "What's basic about basic emotions?", Psychological Review, 97, pp. 315 331.
- Cahn, J.E.: "The Generation of Affect in Synthesized Speech", Journal of the American Voice I/O Society, Volume 8. July 1990. Pages 1-19.
- Carlson, R., Granstrom, B. and Nord, L.: "Experiments with emotive speech-acted utterances and synthesized replicas", Ohala, J. J., Nearey, T. M., Derwing, B. L., Hodge, M. M., & Wiebe, G. E. (Eds.), ICSLP 92 Proceedings (pp. 671-674). University of Alberta, Canada.
- Petrusshin, V.A.: "Emotion in Speech recognition and application to call centers", Proc. Artificial Neural Networks in Engineering (ANNIE99), vol.1, pp. 7-10.
- Elliot, C. and Brzezinski, J. (1998): "Autonomous Agents as Synthetic Characters", AI Magazine, 19:13-30.
- Picard, R.: "Affective computing", The MIT Press, 1997.
- Canh, J.E.: "Generation of Affect in Synthesized Speech", Proceedings of AVIOS'89, Meeting of the American Voice Input/Output Society, 1989.
- Tosa, N. and Nakatsu, R.: "Life-like communication agent - emotion sensing character "MIC" and feeling session character "MUSE"", Proc. of IEEE Conference on Multimedia 1996. 12-19.
- Damasio, A.: "Descarte's Error: Emotion, Reason and the Hitman Brain", London, U.K. Putman, 1994.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N, Votsis, G, Kollias, S., Fellenz, W. and Taylor, J.: "Emotion recognition in human-computer interaction", IEEE Signal Process. Mag., vol.18, no.1, pp. 32-80, 2001.
- Banse, R. and Scherer K.: "Acoustic profiles in vocal emotion expression", J. Personality Social Psych., vol. 70, no. 3, pp. 614-636, 1996.
- Murray, I. and Arnott, J.: "Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", Acoust. Soc. Ante., vol. 93, no. 2, pp. 1097-1108, 1993.
- Devillers L., Chastagnol Ch., Delaborde A., Soury M. and Tahon M.: "Reconnaissance des émotions à partir de l'audio dans le monde reel", Invited lectures, ISIS CNRS 2013.
- Attabi, Y.: "Reconnaissance automatique des émotions à partir du signal acoustique", Montréal. École de technologie supérieure, 2008. 129 p. Maîtrise en génie.
- Xiao, Z.: "Recognition of Emotions in Audio Signals", Doctoral thesis, Information Science, Lyon, France, 2008.
- Vaudable, Ch.: "Analyse et reconnaissance des emotions lors de conversations de centres d'appels", Internal publication. LIMSI, Orsay, France, 2012.

NOTE: Available at https://tel.archives-ouvertes.fr/tel-00758650/en/.

- Chastagnol, C.: "Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes", Université Paris Sud-Paris XI, internal publication, France, 2013.

- Ringeval, F. and Chetouani, M.: "Exploiting a vowel Based approach for acted emotion recognition", A. Esposito et al. (Eds.): HH and HM Interaction 2007, LNAI 5042, pp. 243-254, 2008. Springer-Verlag Berlin Heidelberg, 2008.

- Ringeval, F. and Chetouani, M.: "Une Approche Basée Voyelle pour la Reconnaissance d'Émotions Actées", Institut des Systèmes Intelligents et Robotique, Internal publication, Ivry sur Seine, France, 2008.

- Daassi-Gnaba, H., Zbakh, M. and Lopez Krahe, L.: "Combinaison de reconnaissance de la parole, reconnaissance des émotions et tête parlante codeuse en LPC pour les personnes sourdes et malentendantes", Sciences et Technologies pour le Handicap, Hermès, 2010, 3, pp. 239-253.

- Tahon, M.: "Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot", Internal Publication - doctoral thesis, Université Paris Sud - Paris XI, France, 2012.

NOTE: Available at https://tel.archives-ouvertes.fr/.

- Busso, C., Bulut, M. and Narayanan, S.: "Toward effective automatic recognition systems of emotion in speech. Social emotions in nature and artifact: emotions in human and human-computer interaction", J. Gratch and S. Marsella, Eds., pp. 110-127. Oxford University Press, New York, NY, USA, November 2013.

- Lutfi, S., Montero, J.M., Barra-Chilcote, R., Lucas-Cuesta, J.M. and Gallardo-Antolin, A.: "Expressive speech identifications based on hidden markov model", HEALTHINF 2009 - Proceedings of the 2nd International Conference on Health Informatics. pp. 488-494.

- Bonneau-Maynard, H., Denis, A., Bechet, F., Devillers, L., Lefevre, F., M. Quignard, Sophie Rosset and J. Villaneau.: "Media : évaluation de la compréhension dans les systèmes de dialogue", L'évaluation des technologies de traitement de la langue : les campagnes technolangue (Traité IC2, Série Cognition et Traitement de l'Information), pages 209-231, 2008.

- Devillers, L.: "Les dimensions affectives et sociales dans les interactions humain-robot", Interfaces numériques, 2(1):105-117, 2013.

- Chastagnol, C. and Devillers, L.: "Emotion detection system for human-robot interaction", Colloque du Centre Expertise National en Robotique (CENRob 2013), Evry, France, 04/04 au 05/04 2013.

- Chastagnol, C., Clavel, C., Spurgeon, M. and Devillers, L.: "Designing an emotion detection system for a socially-intelligent human-robot interaction", Towards a Natural Interaction with Robots, Knowbots and Smartphones, Putting Spoken Dialog Systems into Practice. Springer, 2013.

- Chastagnol, C. and Devillers, L.: "Detection d'émotions dans la voix de patients en interaction avec un agent conversationnel animé", Journées d'Etude sur la Parole (JEP 2012), Actes de la conférence conjointe JEP-TALN-RECITAL, page 8p, Grenoble, France, 04/06 au 08/06 2012.

- Tahon, M., Delaborde, A. and Devillers, L.: "Corpus of children voices for mid-level markers and affect bursts analysis", International Conference on Language Resources and Evaluation (LREC 2012), pages 2366-2369, Istanbul, Turkey, 21/05 au 27/05 2012.

- Delaborde, A. and Devillers, L.: "Impact of the social behaviours of the robot on the users emotions: the importance of the task and the subjects age", Workshop Affects, Compagnons Artificiels et Interactions (WACAI 2012), 2012.

- Chastagnol, C. and Devillers, L.: "Collecting spontaneous emotional data for a social assistive robot", 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3 2012), page 5p, 2012.

- Tahon, M., Delaborde, A. and Devillers, L.: "Real-life emotion detection from a speech in human-robot interaction: experiments across diverse corpora with child and adult voices", Annual Conference of the International Speech Communication Association (Interspeech 2011), pp. 3121-3124, Florence, Italy, 2011.

- Delaborde, A., Tahon, M. and Devillers, L.: "Affective links in a child-robot interaction", International Workshop on Emotion: Corpora for Research on Emotion Affect, pp 5, Valetta, Malta, 2010.

- Delaborde, A. and Devillers, L.: "Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers", 3rd ACM Workshop on Affective Interaction in Natural Environments (AFFINE 2010), 2010.
- Delaborde, A., Tahon, M., Barras, C. and Devillers, L.: "A wizard-of-oz game for collecting emotional audio data in a children-robot interaction", AFFINE '09. International Workshop on Affective-Aware Virtual Agents and Social Robots, pp. 3, 2009.
- Soury, M. and Devillers, L.: "Stress detection from audio on multiple window analysis size in a public speaking task", International Conference on Affective Computing and Intelligent Interaction (ACII 2013), pp. 529-533, Geneva, Switzerland, 2013. IEEE Computer Society.
- Tahon, M., Degottex, G. and Devillers, L.: "Usual voice quality features for emotional valence detection", International Conference on Speech Prosody (SP 2012), pp. 4, Shangai, China, 2012.
- Vaudable, Ch. and Devillers, L.: "Negative emotions detection as an indicator of dialogs quality in call centers", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pages 5109-5112, Kyoto, Japan, 25/03 - 30/03 2012.
- Delaborde, A. and Devillers, L.: "Impact of the Social Behaviours of the Robot on the User's Emotions: Importance of the Task and the Subject's Age", Workshop Affects, Compagnons Artificiels et Interactions (WACAI 2012), 2012.
- Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L. and Amir, N.: "Whodunnit searching for the most important feature types signalling emotion-related user states in speech", Computer Speech and Language, 25(1):4-18, 2011.
- Brendel, M., Zaccarelli, R. and Devillers, L.: "Building a system for emotions detection from speech to control an affective avatar", International Conference on Language Resources and Evaluation (LREC 2010), pp. 5, Valetta, Malta, 2010.
- Tahon, M., Delaborde, A., Barras, C. and Devillers, L.: "A corpus for identification of speakers and their emotions", International Workshop on Emotion: Corpora for Research on Emotion Affect, pp. 5, Valetta, Malta, 2010.
- Vaudable, Ch., Rollet, N. and Devillers, L.: "Annotation of affective interaction in real-life dialogs collected in a call-center", International Workshop on Emotion: Corpora for Research on Emotion Affect, pp. 5, Valetta, Malta, 2010.
- Brendel, M., Zaccarelli, R., Schuller, B. and Devillers, L.: "Towards Measuring Similarity Between Emotional Corpora", International Workshop on Emotion: Corpora for Research on Emotion Affect, pp. 5, Valetta, Malta, 2010.
- Brendel, M., Zaccarelli, R., Schuller, B. and Devillers, L.: "A quick sequential forward floating feature selection algorithm for emotion detection from speech", Annual Conference of the International Speech Communication Association (Interspeech 2010), pp. 1157-1160, Makuhari, Chiba, Japan, 2010.
- Devillers, L., Vidrascu, L. and Layachi, O.: "Automatic detection of emotion from vocal expression", A Blueprint for Affective computing: a sourcebook and manual, pp. 232-244. Oxford University Press, 2010.
- Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L. and Ami, N.: "Whodunnit. Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech", Computer Speech and Language, 2010.
- Devillers, L. and Martin, J.-C.: "Emotional corpora", C. Pelachaud, editor, Emotions. John Wiley, 2010.
- Devillers, L. and Martin, J.-C.: "Corpus émotionnels de l'acquisition à la modélisation", C. Pelachaud, editor, Emotions. Hermes, 2010.
- Zara, A., Maffiolo, V., Martin, J.-C. and Devillers, L.: "Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics", ACII, 2007.

- Devillers, L. and Vidrascu, L.: "Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs", International Conference on Speech and Language Processing, pp. 801-804, Pittsburgh, 2006.

NOTE: Available at ftp://tlp.limsi.fr/public/IS061636.PDF.

- Vidrascu, L. and Devillers, L.: "Real-life emotions in naturalistic data recorded in a medical call center", LREC'06 workshop: Emotion, 2006.

- Clavel, C., Vasilescu, I., Devillers, L., Richard, G. and Ehrette, T.: "The SAFE Corpus: illustrating extreme emotions in a dynamic situation", LREC'06 workshop: Emotion, 2006.

- Devillers, L., Vidrascu, L. and Lamel, L.: "Challenges in real-life emotion annotation and machine learning-based detection", Journal of Neural Networks, 18/4, 2005.

- Vidrascu, L. and Devillers, L.: "Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center", ICME, Amsterdam, June 2005.

- Vidrascu, L. and Devillers, L.: "Detection of Real-Life Emotions in Call Centers", InterSpeech, Lisbon, 2005.

NOTE: Available at ftp://tlp.limsi.fr/public/IS052175.PDF.

- Devillers, L., Vasilescu, I. and Lamel, L.: "Annotation and detection of emotion in a task-oriented human-human dialog corpus", ISLE Workshop, Edinburgh, 2002.

NOTE: Available at ftp://tlp.limsi.fr/public/isle02em.pdf.

- Ringeval, F., Chetouani, M. and Schuller, B.: "Novel metrics of speech rhythm for the assessment of emotion", Interspeech-2012, pp. 346-349.

- Brendel, M., Zaccarelli, R. and Devillers, L.: "A quick sequential forward floating feature selection algorithm for emotion detection from speech", Interspeech-2010, pp. 1157-1160.

- Devillers, L., Vaudable, Ch. and Chastagnol, C.: "Real-life emotion-related states detection in call centers: a cross-corpora study", Interspeech, 2010, pp. 2350-2353.

- de Cheveigné, A. and Kawahara, H.: "Comparative evaluation of F_0 estimation algorithms", Eurospeech, Arlborg, Denmark, 2001.

- Clavel, Ch. and Callejas, Z.: "Sentiment analysis: from opinion mining to human-agent interaction", Affective Computing, IEEE Transactions on Affective Computing, vol. PP, no. 99, pp. 1, 2015, doi: 10,1109/TAFFC.2015,2444846, ISSN 1949-3045.

NOTE: Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7122903&tag=1.

- Clavel, C., Vasilescu, I., Devillers, L. and Ehrette, T.: "Fiction database for emotion detection in abnormal situations", Proc. of ICSLP, pp. 2277-2280, Jeju, 2004.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/ICSLP2004Clavel.pdf.

- Clavel, C., Vasilescu, I., Richard, G. and Devillers, L.: "De la construction du corpus émotionnel au système de détection. le point de vue applicatif de la surveillance dans les lieux publics", Revue d'Intelligence Artificielle, numéro spécial Interactions émotionnelles, 20 (4-5): pp. 529-551, 2006.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/RIA_4_Clavel_VF.pdf.

- Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T. and Richard, G.: "Safe corpus : fear-type emotions detection for surveillance application", Proc. of LREC, pp. 1099-1104, Genoa, 2006.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/clavelLREC2006.PDF.

- Clavel, C., Vasilescu, I., Richard, G. and Devillers, L.: "Voiced and unvoiced content of fear-type emotions in the safe corpus", Proc. of Speech Prosody, Dresden, 2006.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/SpeechProsodyClavel.pdf.

- Clavel, C., Vasilescu, I., Devillers, L., Richard, G. and Ehrette, T.: "Fear-type emotions recognition for future audio-based surveillance systems", Speech Communication 50, pp. 487-503, 2008.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/SPECOMClavel.pdf.

- Clavel, C., Vasilescu, I. and Devillers, L.: "Fiction supports for realistic portrayals of fear-type emotional manifestations", Computer Speech and Language, 25(1), pp. 63-83, 2011.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/CSL_ClavelFear.pdf.

- Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., Courcinous, S., Danesi, C., Daquo, A. and Deldossi, M. et al.: "Spontaneous speech and opinion detection: mining call-centre transcripts", Language Resources an Evaluation, vol. 47, number 4, pp. 1089-1125, Springer, 2013.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/LREOpinionMiningClavel.pdf.

- Cornelius, R. R.: The science of emotion: "Research and tradition in the psychology of emotions", NJ, USA: Prentice-Hall, Englewood Cliffs, 1996.

- Cowie, R. and Douglas-Cowie, E.: "Automatic statistical analysis of the signal and prosodic signs of emotion in speech", Proc. Conf. Fourth Int Spoken Language ICSLP 96, vol. 3, 1996, s. 1989-1992.

- Johnson, W. L.; Narayanan, S. S. and Whitney, R.; aj.: "Limited domain synthesis of expressive military speech for animated characters", Proceedings of the 7th International Conference on Spoken Language Processing, Denver, Colorado, USA, 2002, s. 163-166.

- Iida, A.; Campbell, N. and Higuchi, F.; aj.: "A Corpus-based Speech Synthesis System with Emotion", Speech Communication, rocník 40, c. 1-2, 2003: s. 161-187.

- Oudeyer, P.-Y.: "The production and recognition of emotions in speech: features and algorithms", International Journal of Human-Computer Studies, 2003, vol. 59, c. 1-2, pp. 157-183, ISSN 1071-5819, doi: 10,1016/S1071-5819(02)00141-6.

- Tučková, J., Grill, P., Vavřina J. and Bártů, M.: "Speech databases of typical children and children with SLI", LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, 2013.

NOTE: Available at http://hdl.handle.net/11372/LRT-1597.

- Xiao, Z., Dellandrea, E., Dou, W. and Chen, L.: "Multi-stage Classification of emotional speech motivated by a dimensional emotion model", Multimedia Tools and Applications Journal, Springer Netherlands, vol. 46, Nu 1, pp. 119-145, ISSN 1380-7501.

- Shami, M. and Verhelst, W.: "Automatic Classification of Expressiveness in Speech: A Multi-corpus Study", Speaker Classification Ii: Selected Projects, C. Müller, Ed. Lecture Notes In Artificial Intelligence, vol. 4441. Springer-Verlag, Berlin, Heidelberg, 2007, pp. 43-56.

- McGilloway, S., Cowie, R., Cowie, Ed., Gielen, S., Westerdijk, M. and Stroeve, S.: "Approaching automatic recognition of emotion from voice: a rough benchmark", Proceedings of the ISCA Workshop on Speech and Emotion, pp. 207-212, Newcastle, Northern Ireland, 2000.

- Ververidis, D., Kotropoulos, C. and Pitas, I.: "Automatic emotional speech classification", IEEE International Conference on Acoustics, Speech and Signal Processing, 2004, Publisher IEEE, vol. 1, pp. I-593-6.

NOTE: Available at https://scholar.google.gr/scholar?oi=bibs&cluster=2654374610799695O5&btnI=1&hl=en.

- Schuller, B., Rigoll, G. and Lang, M.: "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", Proc. Int. Conf. On Acoustic, Speech and Signal Processing (ICASSP04), vol. 1, pp. 557-560, 2004.

- Petta, P., Cowie, R. and Pelachaud, C: "Emotion-oriented Systems. The Humaine Handbook", Springer-Verlag Berlin Heidelberg 2011, ISBN 11611-2482.

- Nwe, T. L., Foo, S. W. and De Silva, L. C.: "Speech emotion recognition using hidden Markov models", Speech Communication, 2003, vol. 41, pp. 603-623.

- Rabiner, L. R. and Juang, B. H.: "Fundamentals of Speech Recognition", NJ: Prentice-Hall, 1993.
- Womack, B.D. and Hansen, J.H.L.: "Classification of speech under stress using target driven features", Speech Comm. , vol. 20, 1996, pp. 131-150.
- Womack, B.D. and Hansen, J.H.L.: "N-channel hidden Markov models for combined stressed speech classification and recognition", IEEE Trans. Speech Audio Processing, vol.7 (6), 1999, pp. 668-677.
- Scherer K.: "A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology", Proc. 2000 Int. Conf. Spoken Language Processing (ICSLP 2000), Beijing, China, 2000.
- Kienast, M and Sendlmeier, W. F.: "Acoustical analysis of spectral and temporal changes in emotional speech", Proc. ISCA (ITWR) Workshop Speech and Emotion: A conceptual framework for research, Belfast 2000.
- Douglas-Cowie, E., Cowie, R. and Schroder, M.: "A new emotion database: considerations, sources and scope", ITRW on Speech and Emotion, Newcastle, Northem Ireland, UK, 2000.
- Attabi,Y.: "Reconnaissance automatique des émotions à partir du signal acoustique", Montréal. École de technologie supérieure, 2008. Maîtrise en génie.
- Brendel, M., Zaccarelli, R. and Devillers, L.: "A quick sequential forward floating feature selection algorithm for emotion detection from speech", Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), pp. 1157-1160, Makuhari, Chiba, Japan, 2010.
- Brendel, M., Zaccarelli, R. and Devillers, L.: "Building a system for emotions detection from speech to control an affective avatar", International Conference on Language Resources and Evaluation (LREC 2010), pp. 5, Valetta, Malta, 2010.
- Chastagnol, C. and Devillers, L.: "Analysis of Anger across several agent-customer interactions in French call centers", Acoustics, Speech and Signal Processing(ICASSP), 2011 IEEE International French Conference, pp. 4960-4963, 2011.
- Chastagnol, C. and Devillers, L.: "Detection d'émotions dans la voix de patients en interaction avec un agent conversationnel animé", Journées d'Etude sur la Parole (JEP 2012), Actes de la conférence conjointe JEP-TALN-RECITAL, pp. 137-144, Grenoble, France.
- Chastagnol, C., Clavel, C., Courgeon and M., Devillers, L.: "Designing an emotion detection system for a socially-intelligent human-robot interaction", Towards a Natural Interaction with Robots, Knowbots and Smartphones, Putting Spoken Dialog Systems into Practice. Springer, 2013.
- Chastagnol, C.: "Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes", Université Paris Sud - Paris XI, 2013. Français.
- Clavel, C., Vasilescu, I.,Richard, G. and Devillers, L.: "De la construction du corpus émotionnel au système de détection. le point de vue applicatif de la surveillance dans les lieux publics", Revue d'Intelligence Artificielle, numéro spécial Interactions émotionnelles, 20(4- 5):529-551, 2006.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/RIA_4_Clavel_VF.pdf.

- Clavel, C.: "Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales", Télécom ParisTech, 2007.
- Clavel, C., Vasilescu, I., Devillers, L., Richard, G. and Ehrette, T.: "Fear-type emotions recognition for future audio-based surveillance systems", Speech Communication, 50:487-503, 2008.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/SPECOMClavel.pdf.

- Clavel, C., Vasilescu, I. and Devillers, L: "Fiction supports for realistic portrayals of fear-type emotional manifestations", Computer Speech and Language, 25(1):63-83, 2011.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/CSL_ClavelFear.pdf.

- Delaborde, A., Tahon, M. and Devillers, L.: "Affective links in a child-robot interaction", International Workshop on EMOTION: Corpora for Research on Emotion Affect, pp. 75-79, Malta, Valetta. 2010.

- Agnès Delaborde and Laurence Devillers: "Impact du comportement social d'un robot sur les émotions de l'utilisateur : une expérience perceptive", Journées d'Etude sur la Parole (JEP 2012), Actes de la conférence conjointe JEP-TALN-RECITAL, pages 281-288, Grenoble, France, 04/06 au 08/06 2012.

- Laurence Devillers, Ioana Vasilescu and Lori Lamel: "Annotation and detection of emotion in a task-oriented human-human dialog corpus", ISLE Workshop, Edinburgh, 2002.

- Laurence Devillers, Laurence Vidrascu and Lori Lamel: "Challenges in real-life emotion annotation and machine learning-based detection", Journal of Neural Networks, 18/4, 2005.

- Devillers, L. and Vidrascu, L.: "Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs", International Conference on Speech and Language Processing, pp. 801-804, Pittsburgh, 2006.

- Devillers, L. and Vidrascu, L.: "Real-life emotion recognition in speech", Speaker Classification II, Volume 4441 of the series Lecture Notes in Computer Science, pp. 34-42. Springer, 2007.

NOTE: Available at http://link.springer.com/chapter/10.1007/978-3-540-74122-0_4.

- Devillers, L., Vaudable, Ch. and Chastagnol, C.: "Real-life emotion-related states detection in call centers: a cross-corpora study", INTERSPEECH-2010, pp. 2350-2353.

- Devillers, L., Tahon, M., Sehili, M. and Delaborde, A.: "Inference of Human Beings' Emotional States from Speech in Human-Robot Interactions", International Journal of Social Robotics. Audio signal processing - Prediction reliability - Emotion recognition - Human-Robot Interaction. 2014.

- Tahon, M., Delaborde, A., Barras and C., Devillers, L.: "A corpus for identification of speakers and their emotions", International Workshop on EMOTION: Corpora for Research on Emotion Affect, pp. 5, Valetta, Malta, 2010.

- Tahon, M.: "Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot", Université Paris Sud - Paris XI, 2012.

- Tahon, M., Delaborde, A. and Devillers, L.: "Corpus of children voices for mid-level markers and affect bursts analysis", International Conference on Language Resources and Evaluation (LREC 2012), pp. 2366-2369, Istanbul, Turkey, 2012.

- Tahon, M., Delaborde, A. and Devillers, L.: "Real-life emotion detection from speech in human-robot interaction: experiments across diverse corpora with child and adult voices", Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), pp. 3121-3124, Florence, Italy, 2011.

- Vaudable, Ch.: "Analyse et reconnaissance des émotions lors de conversations de centres d'appels", Université Paris Sud - Paris XI, 2012.

- Vidrascu, L. and Devillers, L.: "Detection of Real-Life Emotions in Call Centers", InterSpeech, Lisbon, 2005.

NOTE: Available at ftp://tlp.limsi.fr/public/IS052175.PDF.

- Zhongzhe Xiao: "Recognition of Emotions in Audio Signals", Thèse de doctorat, Ecole central de Lyon, 2008.

- Devillers, L.: "Les dimensions affectives et sociales dans les interactions humain- robot", Interfaces numériques, vol. 2 (1), pp. 105-117, 2013.

- Chastagnol, C. and Devillers, L.: "Emotion detection system for human-robot interaction", Colloque du Centre Expertise National en Robotique, Evry, France, 2013.

- Delaborde, A. and Devillers, L.: "Impact of the social behaviours of the robot on the users emotions: importance of the task and the subjects age", Workshop Affects, Compagnons Artificiels et Interactions, 2012.

- Delaborde, A. and Devillers, L.: "Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers", 3rd ACM Workshop on Affective Interaction in Natural Environments, 2010.

- Vaudable, Ch. and Devillers, L.: "Negative emotions detection as an indicator of dialogs quality in call centers", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 5109-5112, Kyoto, Japan, 2012.

- Devillers, L., Vidrascu, L. and Layachi, O.: "Automatic detection of emotion from vocal expression", A Blueprint for Affective Computing: a sourcebook and manual, pp. 232-244. Oxford University Press, 2010.

- Vidrascu, L., Devillers, L.: "Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center", ICME, Amsterdam, 2005.

- Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., Courcinous, S., Danesi, C., Daquo, A. and Deldossi, M.: "Spontaneous speech, and opinion detection: mining call-centre transcripts", Language Resources an Evaluation, vol. 47, number 4, pp. 1089-1125, Springer, 2013.

NOTE: Available at http://perso.telecom-paristech.fr/~clavel/MaBiblio/LREOpinionMiningClavel.pdf.

- Chul Min Lee, Shrikanth S. Narayanan and Roberto Pieraccini: "Combining acoustic and language information for emotion recognition", 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002.

NOTE: Available at https://www.researchgate.net/publication/221491576_Combining_acoustic_and_language_information_for_emotion_recognition.

- Bjorn Schuller, G. Rigoll and M. Lang: "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", Acoustics, Speech, and Signal Processing. ICASSP-88., 1988 International Conference on 1:I - 577-80, vol. 1, June 2004.

NOTE: Available at https://www.researchgate.net/publication/4087462_Speech_emotion_recognition_combining_acoustic_features_and_linguistic_information_in_a_hybrid_support_vector_machine-belief_network_architecture.

- Laurence Devillers and Ioana Vasilescu: "Prosodic Cues for Emotion Characterization in Real-Life Spoken Dialogs", LIMSI-CNRS, ENST-CNRS TSI, France.

NOTE: Available at https://www.researchgate.net/publication/2876774_Prosodic_Cues_for_Emotion_Characterization_in_Real-Life_Spoken_Dialogs.

- Alexis Clay: "Systèmes interactifs sensibles aux émotions : architecture logicielle", LIPSI-ESTIA/IBRLab-NTHU.

NOTE: Available at http://arxiv.org/pdf/0710.0842.pdf.

- Chloe Clavel: "Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales", LTCI - Laboratoire Traitement et Communication de l'Information.

NOTE: Available at https://tel.archives-ouvertes.fr/pastel-00002533/.

- Kamcke, A.: "Uncertainties Caused by Human Factors - Can an Emotion Detector Help?", IMEKO 2015 World Congress, Prague, 2015.

- Avetisyan, H., Bruna, O. and Holub, J.: "Overview of existing algorithms for emotion classification", IMEKO TC1-TC7-TC13 Joint Symposium, Berkley, 2016.

- Bruna, O., Avetisyan, H. and Holub, J.: "Emotion models for textual emotion classification", IMEKO TC1-TC7-TC13 Joint Symposium, Berkley, 2016.

- Gustafson-Capkova, S.: "Emotions in Speech: Tagset and Acoustic Correlates", 2001.

NOTE: Available at www.speech.kth.se.

- Eckhaus, W.: "The Ginzburg-Landau equation is an attractor", Preprint 746, University Utrecht, 1992.

History

Document History		
V1.1.1	August 2016	Publication