

# Recent Advances in Nonlinear Speech Processing: Directions and Challenges

Anna Esposito, Marcos Faundez-Zanuy, Antonietta M. Esposito,  
Gennaro Cordasco, Thomas Drugman, Jordi Solé-Casals and  
Francesco Carlo Morabito

**Abstract** Humans have very high requirements and expectations when communicating through speech, other than simplicity, flexibility and easiness of interaction. This is because voice interactions do not require cognitive efforts, attention, and memory resources. Voice technologies are however still constrained to use cases and scenarios giving the existing limitations of speech synthesis and recognition systems. Which is the status of nonlinear speech processing techniques and the steps made for cross-fertilization among disciplines? This chapter will provide a short overview trying to answer the above question.

**Keywords** Nonlinear speech processing · Socially believable voice user interfaces · Sound changes · Social and emotional speech features

---

A. Esposito (✉)

Department of Psychology, Seconda Università di Napoli and IIASS, Caserta, Italy  
e-mail: iiass.annaesp@tin.it

M. Faundez-Zanuy

Escola Superior Politècnica Tecnocampus (Pompeu Fabra University), Mataró, Spain  
e-mail: faundez@tecnocampus.cat

A.M. Esposito

Istituto Nazionale di Geofisica e Vulcanologia, sezione di Napoli  
Osservatorio Vesuviano, Rome, Italy  
e-mail: antonietta.esposito@ingv.it

G. Cordasco

Department of Psychology, Seconda Università di Napoli and IIASS, Caserta, Italy  
e-mail: gennaro.cordasc@unina2.it

T. Drugman

University of Mons, TCTS Lab.31, Boulevard Dolez, Mons, Belgium  
e-mail: Thomas.DRUGMAN@umons.ac.be

J. Solé-Casals

Data and Signal Processing Research Group, University of Vic, Barcelona, Spain  
e-mail: jordi.sole@uvic.cat

F.C. Morabito

Università degli Studi “Mediterranea” di Reggio Calabria, Reggio Calabria, Italy  
e-mail: morabito@unirc.it

© Springer International Publishing Switzerland 2016

A. Esposito et al. (eds.), *Recent Advances in Nonlinear Speech Processing*,  
Smart Innovation, Systems and Technologies 48,  
DOI 10.1007/978-3-319-28109-4\_2

# 1 Introduction

Even though contextual instances play a fundamental role in delineating the most appropriate communication tools for implementing successful interactional exchanges [12], nevertheless, spoken messages remain naturally preferred and extremely effective among humans. This is substantiated by the fact that speech based information communication technologies (ICT) are largely accepted and favored among persons. To our knowledge, visual telecommunication tools, such as teleconferencing, are still at an early stage of acceptance, because their “*perceived ease of use (PEOU)*”, and “*perceived usefulness (PU)*”, are strongly affected by both “*individual factors such as anxiety and self-efficacy, and institutional factors such as institutional support and voluntariness*” [26, p.118]. On the contrary, Voice User Interfaces (VUIs), had proven to be largely accepted to the extent that 65+ aged elders are enthusiast to be assisted and monitored for their chronic diseases by a static speaking face [8].

A spoken message produces a precise physical object, a wave of sounds, through which an individual communicates ideas and beliefs, shares knowledge, express needs, feelings, and emotions. The everyday simplicity and flexibility of a such acoustic event in serving as a “container” of countless superimposing and interweaving information, is impressive. The elementary “wave of sounds” will take on several encoding channels, where different streams of data flow together to efficiently build up and successfully shape human exchanges. Among all these encodings, the linguistic code is undoubtedly the most important. It exploits a predefined and shared communication protocol (the language<sup>1</sup>) that allows interactants to decipher a substantial part of the semantic meaning of the delivered message. However, there is a lot of additional information normally sent through speech. Psycholinguistic studies have shown that meanings are conveyed not only by words (intended here as lexicon). During speech production, there exist multiple sets of non-lexical expressions carrying on specific communicative values. Typical non-lexical communicative events at the paralinguistic speech level are, for example, empty and filled pauses signaling, among many other functions, mood states; vocalizations signaling positive or negative feedbacks (“*aah*”, “*hum*”); speech repairs signaling speakers cognitive and emotional states, as well as discourse planning/re-planning strategies; and intonational phrases contour changes allowing to disambiguate meanings [6, 7, 10, 12–14]. The abovementioned speech resources are powerful enough to fulfill plenty of communicative needs without the intervention and independently from the linguistic code, since the process of encoding/decoding for this information is very likely affected by cultural, unconscious, and instinctive communication mechanisms rather than by language production/comprehension rules.

In addition, it is well known that communicative exchanges among humans are not achieved only through speech and linguistic vocal expressions. Written and visual

---

<sup>1</sup>Here “language” is intended to be “the verbal language” as opposed to other general meanings of the term. The interpretation of a “language” as a code can be found in De Saussure [9].

channels, convey linguistic and paralinguistic information that complement or substitute spoken messages and gestures achieve the same pragmatic and semantic speech function [12, 18]. However, at the current technological stage there are few ICT technologies exploiting these channels: speech technologies predominate among all of them and are favorite with respect to visual, graphical and text interfaces. The ultimate speech ICT objectives are guided by the willingness to improve voice services in telecommunication systems, providing a high quality speech synthesis, more efficient speech coding, effective speech recognition, speaker identification, and speaker verification systems in order to significantly spread the VUIs acceptance for information systems such as the mobile Internet (by improving speech synthesis and recognition) and the future generations of wireless communication networks (by improving speech coding).

## 2 Beyond Nonlinear Speech Processing

The nonlinear approach to speech processing had produced advances in several speech engineering fields such as coding, transmission, compression, and synthesis among others, as well as, advances beyond the engineering approach. This is because the functional role of speech, being a human ability, is not constrained to a finite scope and therefore, investigations in one field had produced results in another. Among the topics that had exploited for long time and still exploit nonlinear techniques, it is worth to mention Speech Coding, intended as the ability of an algorithm to code speech in a compact bit-stream such that the amount of transmitted data (the bit rate) would be as low as possible to accommodate transmission channel constraints while preserving speech intelligibility and pleasantness [1, 2, 20]. Low-rate speech coding algorithms have been developed for interactive multimedia services on packet-switched networks such as mobile radio networks, Internet, and mobile network user base, and even more very low bit rate coding at consumer quality will be demanded by the future ICT systems [21, 22, 31].

Two topics of highly nonlinear relevance are Speech Synthesis and Recognition. Humans have very high requirements and expectations when dealing with VUIs, other than simplicity, flexibility and easiness of interaction. This is because voice interactions are an ordinary tool of exchanges among them and do not require, on the user side, cognitive efforts, attention, and memory resources as in the case of graphical and text interfaces. Voice exchanges between humans and machines eliminate delays caused by option menus and can provide very rapidly and complex verbal responses. However current VUIs are not free of constraints. VUIs represent a complex interface option for systems developers since the underlying automated speech recognition (ASR) and Text to Speech (TTS) technology is constrained to context based and speaker dependent applications. Free-form of human-machine conversations are not provided by the current speech technologies. Improvements in dialog management resources are still addressed to specific use scenarios varying from allowing health users to surf the World Wide Web to more complex applications

such monitoring the wellbeing of elderly people, which add, to the complexity of the free-form of conversations also those related to poor speech production (and then more complex efforts for its recognition) because of possible fine motor articulatory impairments due to the age [8, 24, 25, 30]. Current commercial voice enabled systems are Webtalk (<http://www.pcworld.com/article/98603/article.html>) developed by Microsoft, and Siri (<http://www.apple.com/ios/siri/>) developed by Apple. These systems are not free of criticisms and still constrained in the dialogue management to be speaker-dependent, with a restricted dictionary, and favorable environmental conditions. These limitations are mostly due to the many sources of variability affecting the speech signals coarsely grouped by Esposito [16] as: “a) phonetic variability (i.e. the acoustic realizations of phonemes are highly dependent on the context in which they appear), b) within-speaker variability (as result of changes in the speakers physical and emotional state, speaking rate, voice quality), c) across-speaker variability (due to differences in the socio-linguistic background, gender, dialect, and size and shape of the vocal tract), and d) acoustic variability (as result of changes in the environment as well as the position and the characteristics of the transducer)”. Reliable and effective speech recognition and synthesis applications must be able to handle efficiently these variabilities knowing at any stage of the speech recognition/synthesis process which source more than the others is affecting the system efficiency and performance. The general assumption behind these investigations is “that there are rules governing speech variability and such rules can be learned and applied in practical situations” [15, 16]. This point of view is not generally accepted (see [23] for an alternative point of view), since it is related to the classical problem of reconciling the physical and linguistic description of speech, i.e. the invariance issue. Five decades of research in nonlinear speech processing seems to bring convincing arguments on the role of the context (the cultural, organizational, and physical context) in the human communications [12] suggesting to consider the invariance issue context dependent to a certain extent. Two more nonlinear engineering topics such as Voice Analysis, and Conversion (where the quality of the human voice is analysed for clinical and phonetics applications and where techniques for the manipulation of voice characters) produced the flourishing of new speech research fields and new speech applications, such as the analysis of emotional vocal expressions in order to identify speech acoustic emotional features and be able to detect emotional states from speech [3–5, 17, 27, 28] and even more psychopathological disorders such as depression, stress and anxiety [11, 19, 29].

The nonlinear approach to speech processing had gone beyond the acoustic and engineering approach to speech processing, extending its research to the psychological, social, and organizational implications derived from exchanges that are not anymore only among humans, being an automatic system involved. However, in order to be an efficient and effective exchange, the richness of the speech signal must be preserved combining appropriately technological constraints and its social and functional role.

### 3 Contents of this Book

It took over 50 years to realize that speech is beyond speech and therefore nonlinear speech processing should go beyond nonlinear techniques and exploits heuristic and psychological models of human interaction in order to succeed in the implementations of socially believable VUIs and applications for human health and psychological support. This book is signaling advances in these directions taking into account the multifunctional role of speech and what is “outside of the box” (see Björn Schuller’s foreword). To this aim, the book is organized in 6 sections, each collecting a small number of short chapters reporting advances “inside” and “outside” themes related to nonlinear speech research. The themes emphasize theoretical and practical issues for modelling socially believable speech interfaces, ranging from efforts to capture the nature of sound changes in linguistic contexts and the timing nature of speech; labors to identify and detect speech features that help in the diagnosis of psychological and neuronal disease, attempts to improve the effectiveness and performance of Voice User Interfaces, new front-end algorithms for the coding/decoding of effective and computationally efficient acoustic and linguistic speech representations, as well as investigations capturing the social nature of speech in signaling personality traits, emotions and improving human machine interactions.

The coarsely arrangement in 6 scientific sections should be considered only a thematic classification. The sections are closely connected and provide fundamental insights for the cross-fertilization of different disciplines. All the chapters collected in each section are original and never published before. In addition, all the chapters benefited from the live interactions in person among the participants of the successful meeting in Vietri sul Mare under the egide of the 7th biennial international workshop on Non-Linear Speech Processing (NOLISP 2015) which had initiated alternative approaches to speech processing according to the research tradition proposed by the COST Action 277 ([http://www.cost.eu/COST\\_Actions/ict/277](http://www.cost.eu/COST_Actions/ict/277)).

### 4 Conclusions

The readers of this book will get a taste of the major research areas on nonlinear speech processing, different visions on the multifunctional role of speech, different methodologies for analyzing and detecting important speech features, psychological, social, and cognitive disease, and how nonlinear speech processing interact with cognitive and social processes and can shed light on their comprehension and understanding. The research topics proposed by the book are particularly computer science, engineering, signal processing and human-computer interaction oriented and the contributors to this volume are leading authorities in their respective fields. However, interesting psychological, and cognitive aspects are also captured and discussed, letting the book to go, as speech itself, beyond and across scientific disciplines.

## References

1. Arjona Ramírez, M., Minami, M.: Technology and standards for low-bit-rate vocoding methods. In: Bidgoli, H. (ed.) *The Handbook of Computer Networks*, vol. 2, pp. 447–467. Wiley, New York (2011)
2. Arjona Ramírez, M., Minami, M.: Low bit rate speech coding. In: Proakis, J.G. (ed.) *Wiley Encyclopedia of Telecommunications*, vol. 3, pp. 1299–1308. Wiley, New York (2003)
3. Atassi, H., Esposito, A., Smekal, Z.: Analysis of high-level features for vocal emotion recognition. In: *Proceedings of 34th IEEE International Conference on Telecommunication and Signal Processing (TSP)*, pp. 361–366 (2011)
4. Atassi, H., Riviello, M.T., Smekal, Z., Hussain, A., Esposito, A.: Emotional vocal expressions recognition using the cost 2102 italian database of emotional speech. In: Esposito, A., et al. (eds.) *Development of Multimodal Interfaces: Active Listening and Synchrony*, LNCS 5967, pp. 255–267. Springer, Berlin, Heidelberg (2010)
5. Atassi, H., Esposito, A.: Speaker independent approach to the classification of emotional vocal expressions. In: *Proceedings of IEEE Conference on Tools with Artificial Intelligence (ICTAI 2008)*, vol. 1, pp. 487–494 (2008)
6. Butterworth, B.L., Beattie, G.W.: Gestures and silence as indicator of planning in speech. In: Smith, P.T., Campbell, R.N. (eds.) *Recent Advances in the Psychology of Language*, pp. 347–360. Olenum Press, New York (1978)
7. Chafe, W.L.: Cognitive constraint on information flow. In: Tomlin, R. (ed.) *Coherence and Grounding in Discourse*, pp. 20–51. John Benjamins, Amsterdam (1987)
8. Cordasco, G., Esposito, M., Masucci, F., Riviello, M.T., Esposito, A., Chollet, G., Schlögl, S., Milhorat, P., Pelosi, G.: Assessing voice user interfaces: the vAssist system prototype. In: *5th IEEE International Conference on Cognitive InfoCommunications*, pp. 91–96. Vietri sul Mare, 5–7 Nov 2014
9. De Saussure, F.: *Cours de linguistique générale*. Editions Payot, Paris (1922)
10. Esposito, A., Esposito, A.M., Vogel, C.: Needs and challenges in human computer interaction for processing social emotional information. *Pattern Recogn. Lett.* **66**, 41–51 (2015)
11. Esposito, A., Esposito, A.M., Likforman, L., Maldonato, M.N., Vinciarelli, A.: On the significance of speech pauses in depressive disorders: results on read and spontaneous narratives. In this volume (2015)
12. Esposito, A.: The situated multimodal facets of human communication. In: Rojc, M., Campbell, N. (eds.) *Coverbal Synchrony in Human-Machine Interaction*, ch. 7, pp. 173–202. CRC Press, Taylor & Francis Group, Boca Raton, FL (2013)
13. Esposito, A., Marinaro, M.: What pauses can tell us about speech and gesture partnership. In: Esposito, A., et al. (eds.) *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue*. NATO Publishing Series, vol. 18, pp. 45–57. IOS Press, The Netherlands (2007)
14. Esposito, A., Bourbakis, N.G.: The role of timing in speech perception and speech production processes and its effects on language impaired individuals. In: *Proceedings of the 6th International IEEE Symposium on BioInformatics and BioEngineering (BIBE)*, pp. 348–356 (2006)
15. Esposito, A.: The importance of data for training intelligent devices. In: Apolloni, B., Kurfess, C. (eds.) *From Synapses to Rules: Discovering Symbolic Knowledge from Neural Processed Data*, pp. 229–250. Kluwer Academic Press, Dordrecht (2002)
16. Esposito, A.: Approaching speech signal problems: an unifying viewpoint for the speech recognition process. In: Suarez Garcia, S., Baron Fernandez, R. (eds.) *Memoria of Taller Internacional de Tratamiento del Habla, Procesamiento de Vos y el Language*, CIC-IPN Obra Compleata (2000). ISBN: 970-18-4936-1
17. Galanis, D., Karabetos, S., Koutsombogera, M., Papageorgiou, H., Esposito, A., Riviello, M.T.: Classification of emotional speech units in call centre interactions. In: *Proceedings of 4th IEEE International Conference on Cognitive Infocommunications (CogInfoCom2013)*, pp. 403–406. Budapest, Hungary, 2–5 Dec 2013

18. Kendon, A.: *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge (2004)
19. Kiss, G., Tulics, M.G., Sztahó, D., Esposito, A., Vicsi, K.: Language independent detection possibilities of depression by speech. In this volume (2015)
20. Kroon, P.: Evaluation of speech coders. In: Paliwal, K.K., Bastiaan Kleijn, W. (eds.) *Speech Coding and Synthesis*, pp. 467–494. Elsevier Science, Amsterdam (1995)
21. Gibson, J.D.: Speech coding methods, standards, and applications. *IEEE Circuits Syst. Mag.* **5**(4), 30–49 (2005)
22. Faundez-Zanuy, M., Janer, L., Esposito, A., Satue-Villar, A., Roure, J., Espinosa-Duro, V. (eds.): *Nonlinear Analyses and Algorithms for Speech Processing*, LNAI 3817. Springer, Berlin, Heidelberg (2006)
23. Lindblom, B.: Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W., Marchal, A. (eds.) *Speech Production and Speech Modeling*, pp. 403–439. Kluwer, Dordrecht (1990)
24. Meena, R., Skantze, G., Gustafson, J.: Data-driven models for timing feedback responses in a map task dialogue system. *Comput. Speech Lang.* **28**, 903–922 (2014)
25. Milhorat, P., Schlögl, S., Chollet, G., Boudy, J., Esposito, A., Pelosi, G.: Building the next generation of personal digital assistants. In: *Proceedings of 1st IEEE International Conference on Advanced Technologies for Signal and Image Processing–ATSIP’2014*, pp. 458–463. Sousse, Tunisia, 17–19 Mar 2014. ISSN 978-1-4799-4888-8/14/
26. Park, N., Rhoads, M., Hou, J., Lee, K.M.: Understanding the acceptance of teleconferencing systems among employees: an extension of the technology acceptance model. *Comput. Hum. Behav.* **39**, 118–127 (2014)
27. Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.P., Ebrahimi, T., Lalanne, D., Schuller, B.: Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recogn. Lett.* Elsevier (2014)
28. Schullerm, B.: Deep learning our everyday emotions: a short overview. In: Bassis et al. (eds.) *Advances in Neural Networks: Computational and Theoretical Issues*. Series: SIST Series, vol. 37, pp. 339–346. Springer, Berlin, Heidelberg (2015)
29. Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., Morency, L.P.: Automatic audio-visual behaviour descriptors for psychological disorder analysis. *Special Issue on Best of Face and Gesture 2013: Image Vis. Comput.* **32**(10), 648–658 (2014)
30. Skantze, G., Hjalmarsson, A.: Towards incremental speech generation in conversational systems. *Comput. Speech Lang.* **27**, 243–262 (2013)
31. Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.): *Progress in Nonlinear Speech Processing*, LNCS 4391. Springer, Berlin, Heidelberg (2007)

Recent Advances in Nonlinear Speech Processing

Esposito, A.; Faundez-Zanuy, M.; Esposito, A.M.;

Cordasco, G.; Drugman, Th.; Solé-Casals, J.; Morabito,

F.C. (Eds.)

2016, XIV, 294 p. 84 illus., 62 illus. in color., Hardcover

ISBN: 978-3-319-28107-0