Towards Hierarchical Cognitive Systems for Intelligent Signal Processing

Rüdiger Hoffmann[1] and Matthias Wolff[2]

[1] Technische Universität Dresden,
Professur Systemtheorie und Sprachtechnologie, 01062 Dresden, Germany
ruediger.hoffmann@tu-dresden.de
http://www.ias.et.tu.dresden.de
[2] Brandenburgische Technische Universität Cottbus,
Lehrstuhl Kommunikationstechnik, 03046 Cottbus, Germany
matthias.wolff@tu-cottbus.de
http://www.tu-cottbus.de/kommunikationstechnik/

Abstract. Speech and many other acoustic signals show a hierarchical structure which has to be considered if systems for speech and audio processing are developed. Because systems for speech recognition and for speech synthesis follow the same hierarchy, a unified approach (UASR) was proposed in the year 2000, which was implemented during the following decade. The general application of Finite State Transducers (FST) results in a very efficient technology at all symbolic levels of the hierarchy. UASR proved to be successful not only for speech processing but also for many applications to other biological, technical, or musical signals, resp. At the same time, the idea of cognitive dynamic systems became popular mainly due to the work of S. Haykin. It is very promising to expand the UASR system to a hierarchical cognitive dynamic system, combining the hierarchical structure of UASR with the approach of cognitive systems, which is mainly elaborated for the so-called cognitive radio so far. The target system, the structure of which was defined now, will perform intelligent processing of speech and other signals.

Keywords: intelligent signal processing, hierarchical systems, cognitive systems, acoustic pattern recognition, audio processing

1 Introduction

Speech processing systems are basing on two basic principles which were formulated, e. g., in one of the fundamental papers on speech recognition [1]: "The first fundamental hypothesis is based on consistent evidence [...] indicating that the primary information-bearing attitude of the speech signal is the temporal variation of the short duration amplitude spectrum. [...] The second foundational rule, which is borne out by the results of [...] psychophysical experiments [...], asserts that speech is a composite signal, hierarchically organized so that simpler patterns at one level are combined in a well-defined manner to form more complex patterns at the succeeding level. [...] The structures at each level

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

of the hierarchy serve to constrain the ways in which the individual patterns associated with that level can be combined."

The growing success of statistical approaches in speech technology during the 1990-th resulted in a convergence of speech recognition and speech synthesis which had developed hitherto in separate ways. This was mainly due to the necessity of large databases or knowledge sources in both branches. This development had been predicted, e. g., in a classical textbook [2]: "Advanced systems both for synthesis and for recognition need the same speech knowledge, and there is considerable advantage for the two applications to be studied together. [. . .] I predict that the most significant progress in the more advanced forms of speech synthesis and recognition will in future come from research teams with a strong interest in both problems." The development of the so-called HMM synthesis was the most important result of this generalized sight.

Both approaches, which have been addressed by the citations above, the hierarchical as well as the analysis-synthesis concept, have been united in our UASR (Unified Approach for Speech Synthesis and Recognition) which was initiated as a long-term project in the year 2000 [3]. The following is an extended abstract of an overview of our work through the last decade, which is mainly intended to provide a commented collection of references to the original work.

2 The UASR System

The implementation of the aforementioned ideas resulted in a system which was composed from three components, each of which offering a hierarchical structure:

- the analysis (or recognition) branch with information flowing bottom-up,
- the synthesis branch with information flowing top-down,
- a set of databases, one for each hierarchy level, which are accessed by the analysis as well as the synthesis component of the respective level.

The first implementation included the feature, phonetic, lexical, and syntactic levels. It was described in [4], [5], and [6].

The algorithmic design of the processing elements at the "symbolic" hierarchy levels proved to be a challenge. UASR is a big software systems, and it was necessary to look for a method to design the components in an uniform and economic way. The application of finite state transducers which were introduced in the speech technology some years before by M Mohri [7], turned out as the most efficient way. After a general redesign, UASR is now implemented in FST technology completely [8].

We want to stress that UASR is not only a powerful experimental platform. Additionally, we implemented an embedded version in cooperation with the Fraunhofer society for speech control applications [9] [10]. The recent version works basing on an FPGA, but a special processor will be the final goal.

3 Useful Non-speech Applications

There are many applications of pattern recognition algorithms to acoustical and other signals from the real world. In the past, numerical pattern recognition methods proved to be sufficient for solving those tasks. Structural pattern recognition was essentially restricted to applications with speech signals. This situation changed around the year 2000 when the performance of classical approaches was not satisfactory for new tasks in non-destructive testing and other branches. Using UASR, we were able to demonstrate in may cases that structural methods are able to introduce much progress to different classes of signals. We mention some examples as follows:

- **Non-destructive testing.** This was the dominant application area, supported by a cooperation with the Fraunhofer society. The results are summarized in [11], [12], and the thesis [13].
- **State monitoring of machines.** This tasks is similar to the aforementioned, but does not require a sharp decision. Instead, the degree of membership to the one and only defined class has to be calculated. Different projects which were performed are summarized in [14].
- **Biological signals.** From different examples, we want to mention here the measurement of the blood pressure under real conditions (i. e., a moving test person), because this was a real challenging task [15].
- **Musical instruments.** This is a specific version of non-destructive testing which we are dealing with since many years in cooperation with the industry [16]. Among other results, we have shown that it is possible to identify specific instruments by structural methods [17] [18].
- **Musical signals.** The investigation of musical signals is useful for a number of applications in music retrieving, measuring similarities between musical works, etc. [19]. Additionally, speech research may benefit from this because basic findings about rhythm may be transferred to speech signals, which is an very actual problem in prosody research [20].

A more complete overview is given in [21] and [8].

4 Towards Cognitive Systems

Some years ago, S.Haykin coined the term "cognitive dynamic systems" for systems which show a purposeful behavior like human beings [22]. They are able to develop an internal model of their environment and, basing on this, to influence their environment actively. Surprisingly, elaborated applications of this theory are existing not only in the traditional fields of artificial intelligence (including speech technology), but mainly in "cognitive signal processing systems" like the cognitive radar and the cognitive radio [23].

Systems like cognitive radio are not really hierarchical (although the application of the well-known OSI model introduces a hierarchy). Therefore there is

much similarity between non-hierarchic cognitive systems and the classical theory of automatic control. Haykin points out, however, that cognitive systems in biology are organized in a hierarchical way, demonstrating it with a prominent example from [24]. The block diagram of an hierarchical perception-action system of this kind is very similar to that of UASR.

Special attention has to be directed to the interaction between the levels of the hierarchy. Although the information flow of the analysis branch is bottom-up, and of the synthesis branch top-down, the performance will be improved by bidirectional interaction. The cortical algorithm which was originally developed for modeling the visual cortex [25], seems to be very promising for the implementation in UASR [26] [27].

5 Conclusion

Basing on the ideas discussed above, UASR offers some potential directions for further development, as follows:

- The success of applying UASR to many non-speech problems resulted in the idea to generalize the hierarchical analysis-synthesis systems towards an intelligent system for hierarchical signal processing. It is discussed in detail in [8] and as an overview in [28].
- The step from the existing UASR towards a real "cognitive" system will be done by adding a cognitive backend at the top of the speech-related hierarchy. This could be performed in various ways, e. g. by adding a translation component like in the former Verbmobil system [29]. We decided to follow a solution which was proposed in [30], aiming to a dialogue controller. For a flowgraph of the extended UASR, cf. [8] or [31]. It is a challenge to demonstrate that the uniform application of FST technology (maybe generalized to a theory of Petri net transducers) can be extended to the semantic components also [32].
- The principle of "analysis by synthesis" (AbS), which plays an essential role in the development of speech technology, was successfully applied in introducing new approaches for the parametric speech synthesis (so-called HMM synthesis [33] [34]). We expect some progress in the personalization of speech technology by applying this principle to deeper studies of human prosody [35].

Acknowledgments. This work was partially funded by

- the Deutsche Forschungsgemeinschaft (DFG) under grants Ho 1674/3, Ho 1674/7, and Ho1684/8,
- the German Federal Ministry of Education and Research (BMBF), and
- the Arbeitsgemeinschaft industrieller Forschungsvereinigungen "Otto von Guericke" (AiF).

References

1. S. E. Levinson: Structural methods in automatic speech recognition. Proceedings of the IEEE 73 (1985) 11, 1625–1650.
2. J. N. Holmes: Speech Synthesis and Recognition. London: Van Norstrand Reinhold 1988.
3. M. Eichner; M. Wolff; R. Hoffmann: A unified approach for speech synthesis and speech recognition using stochastic Markov graphs. Proc. ICSLP, Beijing 2000, vol. 1, 701–704.
4. S. Werner, M. Eichner, M. Wolff, R. Hoffmann: Towards spontaneous speech synthesis – Utilizing language model information in TTS. IEEE Trans. on Speech and Audio Processing 12 (2004) 4, 436–445.
5. M. Eichner: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen – Akustische Analyse und Modellierung. Dresden: TUDpress 2007 (Studientexte zur Sprachkommunikation, vol. 43).
6. S. Werner: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen – Sprachmodell und Aussprachemodellierung. Dresden: TUDpress 2008 (Studientexte zur Sprachkommunikation, vol. 48).
7. M. Mohri: Finite-state transducers in language and speech processing. Computational Linguistics 23 (1997) 2, 269–311.
8. M. Wolff: Akustische Mustererkennung. Dresden: TUDpress 2011 (Studientexte zur Sprachkommunikation, vol. 57).
9. G. Strecha, M. Wolff, F. Duckhorn, S. Wittenberg, C. Tschöpe: The HMM synthesis algorithm of an embedded unified speech recognizer and synthesizer. Proc. Interspeech, Brighton 2009, 1763–1766.
10. G. Strecha, M. Wolff: Speech synthesis using HMM based diphone inventory encoding for low-resource devices. Proc. IEEE ICASSP, Prague 2011, 5380–5383.
11. C. Tschöpe, D. Hentschel, M. Wolff, M. Eichner, R. Hoffmann: Classification of non-speech acoustic signals using structure models. Proc. IEEE ICASSP, Montreal 2004, vol. 5, 653–656.
12. C. Tschöpe, M. Wolff: Statistical classifiers for structural health monitoring. IEEE Sensors Journal 9 (2009) 11, 1567–1576.
13. C. Tschöpe: Akustische zerstrungsfreie Prfung mit Hidden-Markov-Modellen. Dresden: TUDpress 2012 (Studientexte zur Sprachkommunikation, vol. 60).
14. S. Wittenberg: Statistische Ein-Klassen-Signalbewertung mit akustischen Datenbasen selbstbeschreibender Daten. Dresden: TUDpress 2012 (Studientexte zur Sprachkommunikation, vol. 63).
15. M. Wolff, U. Kordon, H. Hussein, M. Eichner, C. Tschöpe; R. Hoffmann: Auscultatory blood pressure measurement using HMMs. Proc. IEEE ICASSP, Honolulu 2007, vol. 1, 405–408.
16. G. Ziegenhals: Subjektive und objektive Beurteilung von Musikinstrumenten. Dresden: TUDpress 2010 (Studientexte zur Sprachkommunikation, vol. 51).
17. M. Eichner, M. Wolff, R. Hoffmann: Instrument classification using hidden Markov models. Int. Conf. on Music Information Retrieval (ISMIR), Victoria 2006, 349–350.
18. M. Eichner, M. Wolff, R. Hoffmann: An HMM based investigation of differences between musical instrumentsof the same type. Proc. Int. Congress of Acoustics (ICA), Madrid 2007, 5 pp.
19. S. Hübler, R. Hoffmann: Evaluation of onset detection algorithms in popular polyphonic music on a large scale database. Proc. of the 130th Audio Engineering Society Convention, London 2011, 1265–1270.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

20. S. Hübler, R. Hoffmann: Comparing music and speech with a closer look on automatic music information retrieval. In: A. Esposito et al. (eds.): Towards Autonomous, Adaptive, and Context-aware Multimodal Interfaces. Berlin etc.: Springer 2011 (Lecture Notes in Computer Science, vol. 6456), 376–386.
21. R. Hoffmann, M. Eichner, M. Wolff: Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: A. Esposito et al. (eds.): Verbal and Nonverbal Communication Behaviours. Berlin etc.: Springer 2007 (Lecture Notes in Artificial Intelligence, vol. 4775), 200–218.
22. S. Haykin: Cognitive dynamic systems. Proc. IEEE ICASSP, Honolulu 2007, vol. 4, 1369–1372.
23. S. Haykin: Cognitive Dynamic Systems. Perception-action Cycle, Radar and Radio. Cambridge University Press 2012.
24. J. M. Fuster: Cortex and Mind – Unifying Cognition. New York: Oxford University Press 2003.
25. T. S. Lee, D. Mumford: Hierarchical Bayesian inference in the visual cortex. JOSA 20 (2003) No. 7.
26. R. Römer; T. Herbig: Konzeptionelle Beschreibung des cortikalen Algorithmus und seine Anwendung in der automatischen Sprachverarbeitung. In: R. Hoffmann (ed.): Elektronische Sprachsignalverarbeitung, Dresden 2009, Bd. 1 (Studientexte zur Sprachkommunikation, vol. 53), 33–40.
27. R. Römer: Untersuchungen zum cortikalen Algorithmus unter Verwendung von bidirektionalen HMMs. In: M. Wolff (ed.): Elektronische Sprachsignalverarbeitung, Cottbus 2012 (Studientexte zur Sprachkommunikation, vol. 64), 252–261.
28. M. Wolff, R. Hoffmann: An approach to intelligent signal processing. In: A. Esposito et al. (eds.): Cognitive Behavioural Systems. COST 2102 International Training School, Dresden 2011, Revised Selected Papers. Berlin etc.: Springer 2012 (Lecture Notes in Computer Science, vol. 7403).
29. W. Wahlster: Verbmobil – Foundations of Speech-to-Speech Translation. Berlin etc.: Springer 2000.
30. M. Huber, C. Kölbl, R. Lorenz, R. Römer, G. Wirsching: Semantische Dialogmodellierung mit gewichteten Merkmal-Werte-Relationen. In: R. Hoffmann (ed.): Elektronische Sprachsignalverarbeitung, Dresden 2009, vol. 1 (Studientexte zur Sprachkommunikation, vol. 53), 25–32.
31. M. Wolff, R. Römer, R. Hoffmann: Hierarchische kognitive dynamische Systeme zur Sprach- und Signalverarbeitung. In: M. Wolff (ed.): Elektronische Sprachsignalverarbeitung, Cottbus 2012 (Studientexte zur Sprachkommunikation, vol. 64), 159–178.
32. M. Huber, R. Lorenz: Petri net transducers in semantic dialogue modelling. In: M. Wolff (ed.): Elektronische Sprachsignalverarbeitung, Cottbus 2012 (Studientexte zur Sprachkommunikation, vol. 64), 286–297.
33. A. Falaschi, M. Giustiniani, M. Verola: A hidden Markov model approach to speech synthesis. Proc. Eurospeech, Paris 1989, 187–190.
34. K. Tokuda et al.: Speech parameter generation algorithms for HMM-based speech synthesis. Proc. IEEE ICASSP, Istanbul 2000, 1315–1318.
35. R. Hoffmann: Analysis-by-synthesis in prosody research. Keynote, Proc. 6th International Conference on Speech Prosody, Shanghai 2012, 1–6.