Analysis-by-Synthesis in Prosody Research

Rüdiger Hoffmann

Chair for System Theory and Speech Technology, Technische Universität Dresden, Germany

Ruediger.Hoffmann@tu-dresden.de

Abstract

It was early recognized in the history of speech technology, that prosody plays an essential role in the communication process and that it is therefore necessary to include prosodic components into the speech-based systems for man-computer interaction. Recent text-to-speech (TTS) systems show prosodic components at an elementary level (intonation and duration) for good comprehensibility, but it is also obvious that these components are not powerful enough to produce speech with high naturalness and personality. On the other hand, systems for automatic speech recognition (ASR) consider the prosody more or less implicitly, and we have only few examples where prosodic features are explicitly used for improving the recognition results. This talk is an attempt to give a more general view on the inclusion of prosody in speech technology. During the last decade, reconsidering the paradigm of analysis-by-synthesis (AbS) in speech technology has produced some algorithmic progress in TTS and in ASR as well. The system UASR (Unified Approach for Speech Synthesis and Recognition) of the TU Dresden was designed to demonstrate the AbS approach in a hierarchical way. It is now time to discuss how prosodic components could be included in such systems. The inclusion of rhythmic phenomena seems to be the most difficult but also very promising subtask. Possibly speech processing can benefit from musical signal processing where the identification of rhythm is a very natural task.

Index Terms: history of speech technology, Analysis-by-Synthesis, UASR, cognitive systems, hierarchical systems, speech dialogue systems

1. Introduction

Prosody research is growing very much during the last years. This is mainly due to the growing interest in social interaction where speech communication establishes only one of the communication modes. We have learned that speech prosody is not only part of linguistics, but also forms a bridge to non-linguistic communication and, above all, non-verbal modes like gesturing. Speech technology has utilized the progress in prosody research in a limited way until today. This is true for speech synthesis, but even more for speech recognition. Speech technology is now advancing towards speech dialogue systems. It seems to be useful, to reconsider the inclusion of prosody in such systems from an engineering point of view.

The investigation of prosodic effects in engineering has its own history. Roughly speaking, it started with a kind of trial and error, which was more and more refined to that epistemological approach which we call now Analysis-by-Synthesis (AbS).

AbS is very natural in speech technology because everybody will agree that building a speech based system means to design and implement a model of that what humans do if they are speaking or listening. AbS allows to optimize the modeling process (Figure 1) to achieve maximal similarity between the biological system and its engineering counterpart.

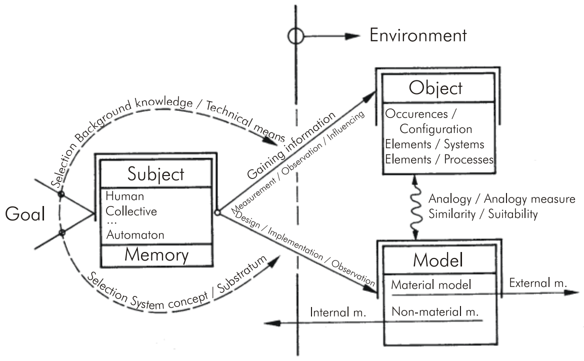


Figure 1: *Model method (adapted from* [1]*).*

2. AbS in the history of prosody research

2.1. The pre-electronic era [2]

It is interesting to note that Wolfgang von Kempelen, the forefather of the modern speech synthesis, recognized the importance of the speech melody for his speaking machine: “Ich habe oft nachgedacht, ob man nicht […] dahin kommen könnte […], dieses Fallen und Steigen des Tones nach Willkühr zu bewirken und dadurch […] wenigstens eine Abwechslung der Stimme bey dem Sprechen zu erhalten, welches meiner Maschine, die dermalen alles in einem Tone fortspricht, erst die rechte Annehmlichkeit geben würde.“ [3, p. 413]. He describes first attempts with a manual control.

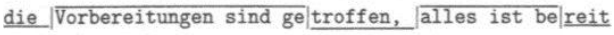
One century later, the special interest of the experimental phonetics in measuring the pitch contour as one of the most important physical phenomena of the prosody was activated because many foreign languages (the “colonial languages”) had to be investigated. The analysis was performed mainly by interpreting the recordings of kymographs or phonographs. This very complicated and time-consuming process used a number of tools which we have described in [4]. Of course, there was no possibility to verify the results by means of re-synthesis.

2.2. The vocoder era

There were different attempts in speech synthesis at the beginning of the electronic era. The real breakthrough was achieved with the invention of the channel vocoder by K. O. Schmidt [5] and H. W. Dudley [6]. The subdivision of the device in an analyzer and a synthesizer enabled an analysis-by-synthesis process in a very effective way [7]. The existence of a separate channel for the fundamental frequency allowed the demonstration of the effect of pitch manipulation and thus the experimental investigation of prosodic contours. Some sound examples from the early vocoders are still available.

The analysis-by-synthesis activities in speech prosody go back to vocoder experiments. The linguists A. V. Isačenko (1910 - 1977, a well-known slavist) and H.-J. Schädlich

(* 1935, later known as a novelist) were among the first who developed models for the quantitative description of prosodic effects [8]. The English translation of their report [9] includes a disk with some of the test sentences. This test material consists of German sentences with a fundamental frequency which was manipulated to have only two values, e. g. [8]:


die |Vorbereitungen sind ge|troffen, |alles ist be|reit

Experiments showed that there is still enough prosodic information to recognize the correct grammatical structure of the sentences. The manipulation was performed using the Dresden vocoder with support of W. Tscheschner and later with the Ericsson vocoder, supported by G. Fant.

2.3. Prosodic experiments with formant synthesizers

The first channel vocoders have been large and expensive. There was some doubt whether they could be widely used in commercial applications. Also, the speech signal had “inhuman” quality and limited comprehensibility. It became clear that there are more effective kinds of parameterization of the speech signal, and other vocoder types than the channel vocoder arose. Formant coding proved to be a very effective approach

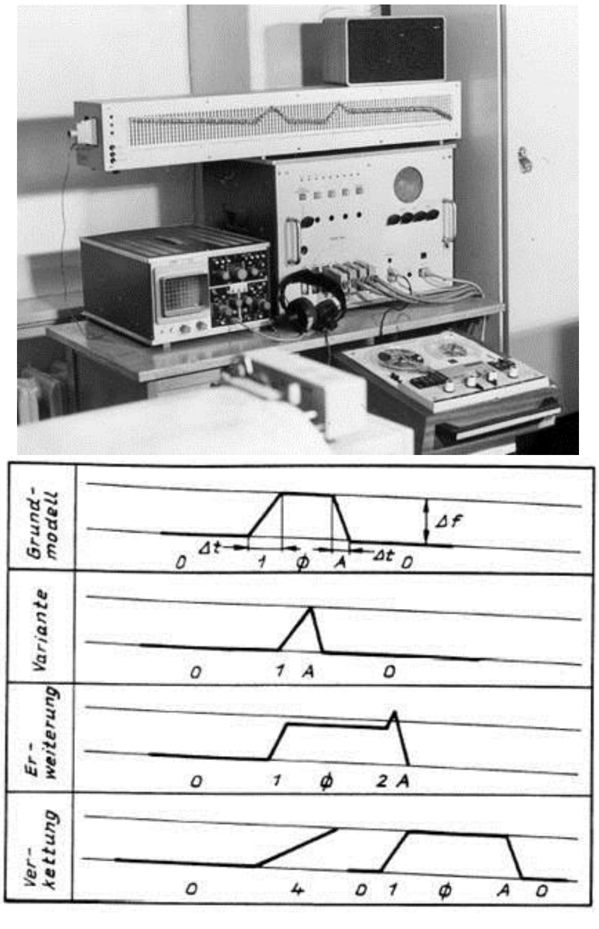


Figure 2 – *Prosodic experiments with ROSY 4200. Above: Experimental setup with the synthesizer terminal ROSY (middle right) and the contour generator (above). The control computer is not shown. – Below: Models of suprasegmental fundamental frequency contours from* [12].

Consequently, the early types of speech synthesis terminals also followed the principle of formant synthesis. This development was strongly influenced by the work of G. Fant and can be illustrated using the history at different places. We have described this way of early speech synthesis especially at the TU Dresden under the guidance of W. Tscheschner (1927 - 2004) in [10]. The prosodic investigations were connected to the ROSY project of the 1970-th. ROSY was a process computer controlled four-formant speech synthesizer. A small series of the synthesizers was produced by the Dresden computer company Robotron where the name of the device comes from (RObotron SYnthesizer). Formant synthesizers are very well suited for prosodic experiments (and even for singing) due to the presence of a separate excitation generator with controllable pitch.

The prosody research for the speech synthesizers of the TU Dresden was performed in close cooperation with the Humboldt University at Berlin. It can be divided in two phases. In the first one, the microintonation at the sound transitions of German was investigated using natural speech material. Different types of transitions were classified, and a group of five was finally proposed for the application in speech synthesis [11]. They were implemented in the hardware of the ROSY synthesizer.

In the second phase, analysis-by-synthesis experiments on the German macrointonation had been performed [12] with synthetic speech. For this purpose, the synthesis terminal ROSY was complemented by a contour generator which allowed influencing the intonation of the synthesizer by hardware. Basing on listening experiments, a number of standard contours could be proposed for the speech synthesis (Figure 2). Some examples of the test sentences in different intonation versions (monotonous / linear declination / declination plus accentuation) are still available as audio files.

2.4. Prosody in concatenative speech synthesis

The idea to synthesize natural sounding speech by concatenating speech segments from a database with real speech is not really new. With the invention of the magnetic storage of audio signals, the idea of the so-called concatenative synthesis emerged. The “digital” renaissance of the idea came with the availability of powerful PCs at the beginning of the 1990-th. They offered enough memory for the speech samples as well as enough computing power for the text and signal processing of the complete text-to-speech conversion chain. Unfortunately, prosodic manipulations were now more challenging compared to formant synthesizers. The TD-PSOLA algorithm [13] was the predominant solution und paved the way to a broad application of speech synthesis in time domain.

The emerging TTS technology required reliable control of the prosodic parameters for whole sentences or phrases. Therefore, quantitative models of macrointonation received more and more attention. A real breakthrough was achieved by the model of H. Fujisaki (e. g., [14]) which was applied successfully to many languages. Much effort was made to find effective training algorithms for the parameters of the Fujisaki model (e.g., [15]).

A systematical investigation of the German prosody was performed with the MFGI (“Mixdorff Fujisaki German Intonation”) model. In this framework, we compared the prosodic quality of concatenative TTS for different prosody models and found that MFGI performed favorable [16].

3. From AbS to cognitive systems

3.1. The UASR platform as a prototype

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 in a classical textbook [17]: "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 [18, 19].

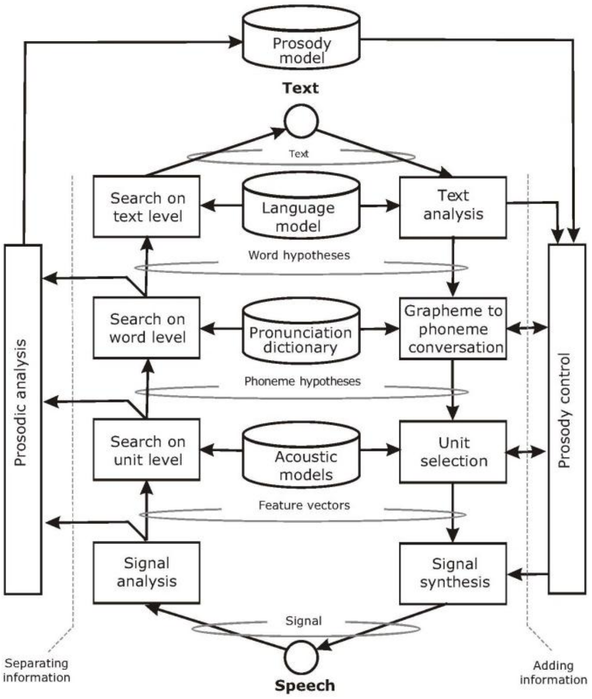


Figure 3: *Unified approach for speech recognition and synthesis (UASR), as it was proposed in the year 2000* [20].

Recognizing these tendencies, we started around 2000 the development of a prototype system called UASR which integrated the elements of a typical speech recognizer and a speech synthesizer with common databases (Fig. 3). This system was implemented over the decade past 2000. About the progress, cf. [20 - 23]. The aims of the project have been:

- improved understanding of the algorithms by means of the principle of AbS in an *hierarchical* system,
- improved understanding of the reasons why speech recognition results are erroneous,
- development of components for parametric synthesis basing on statistically trained models,
- building a toolbox for practical (also embedded [24]) applications of speech recognition and synthesis,
- building the baseline system for numerous applications in the field of non-speech signals like biological [25], technical [26], or environmental signals, and music.

3.2. Cognitive dynamical systems

UASR is a prototype for a very up-to-date research field. S. Haykin coined the term *cognitive dynamic systems* for systems which show a purposeful behavior like human beings [27]. They are able to develop an internal model of their environment and, basing on this, to influence their environment actively. Obviously, there are close connections to the classical theory of automatic control (Fig. 4).

Surprisingly, elaborated applications of this theory are existing not only in the traditional fields of artificial intelligence (including speech technology), but also in "cognitive signal processing systems" like the cognitive radar [28] and the cognitive radio [29].

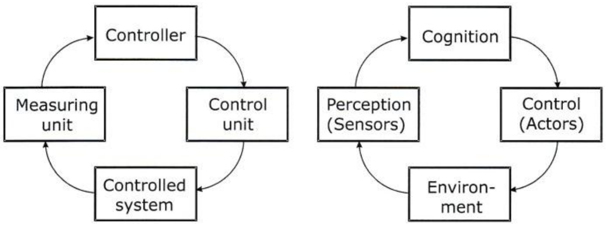


Figure 4: *Cognitive dynamic system as proposed by Haykin (right), compared with the loop of a classical system for automatic control (left).*

Considering the human as the most universal cognitive dynamic system, we must take into account the hierarchical structure of the information processing which is obviously essential for its function [30]. At all levels of the hierarchy, a combination of abstraction (bottom-up) and prediction (top-down) occurs. Therefore we find technical systems, which have (at least in a rudimentary way) a comparable hierarchical structure, mainly in the field of man-machine interaction, e. g. in processing speech, images, gestures, etc. This explains the formal analogy of the biological findings with the UASR structure in Fig. 3.

Other existing cognitive systems like cognitive radio require this hierarchical structure in less extent. A formal similarity exists due to the application of the well-known OSI reference model (Open Systems Interconnection Reference Model) which leads to the inclusion of the same statistical learning and decision algorithms.

Coming back to speech technology and AbS, it seems to be a natural extension of the UASR approach to form a cognitive system by adding a "speech understanding" component which acts as the cognitive module in the sense of Fig. 4. This could be very useful because it is generally recognized that the existing recognizers and synthesizers suffer from the lack of a real "understanding" of that what they do.

On a second glance, the problem arises how to design the interface between UASR and the understanding component. Speech understanding is a task of computer linguistics which normally expects an input of formally correct texts. This input, however, cannot be delivered by a system which processes *spoken* language due to two reasons. At first, natural (spontaneous) speech is not regular in a strong sense. Secondly, we know that a speech recognizer makes errors.

These problems had been already considered in a former big research project, called Verbmobil, where the understanding component was a translation software [31]. We can apply experiences from this project if we are enlarging the UASR structure.

3.3. An UASR based hierarchical cognitive system

Considering the arguments from the previous section, an extension of the UASR approach to a cognitive system could look like Fig. 5. The cognitive backend could be implemented in several ways, e. g. as a speech translation system similar to Verbmobil [31] or a dialogue system which can answer inquiries or control a technical system. In our case, it is intended to build a dialogue system basing on the concept in [32]. (For a flowchart, cf. [34, p. 252].) In all cases, the restriction on a certain domain will be necessary to keep the computing and database expenses in reasonable limits.

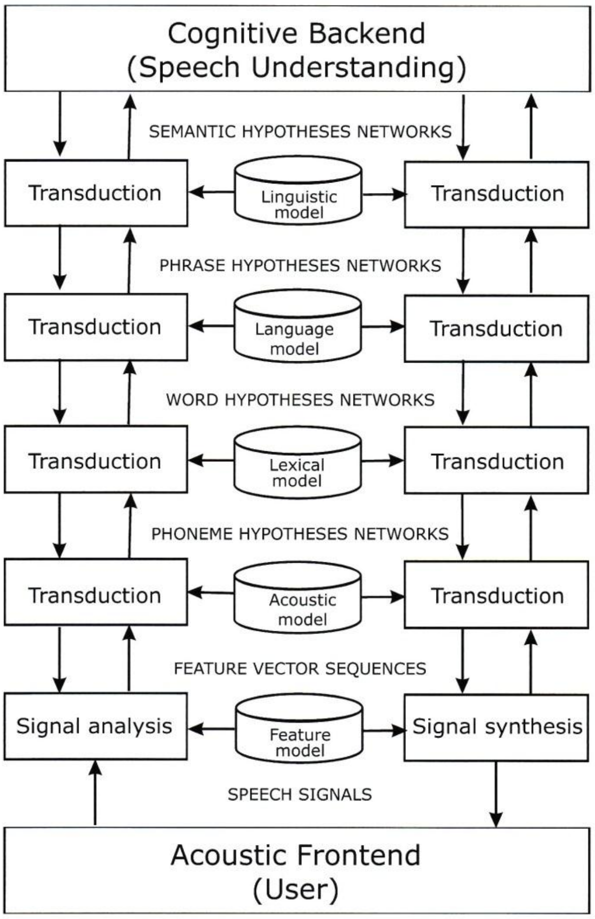


Figure 5: *Extension of UASR to a hierarchical cognitive dynamic system.*

There are two important differences between Fig. 3 and Fig. 5. At first, the specific processing modules in Fig. 3 are replaced by uniform “transduction” modules above the signal processing level. Indeed, the algorithms in speech technology could be standardized in large extent by applying Finite State Transducers [33]. UASR is now completely basing on FST technology [34]. It was shown [35] that also the semantic components can be formalized within a FST framework.

As a second difference, the pure bottom-up structure of the (left) analysis branch and the pure top-down structure of the (right) synthesis branch of UASR are supplemented in Fig. 5 by connections in the opposite direction. This allows for a better consideration of the interplay between abstraction and prediction. A suited approach for this improved interaction between the levels is given in [36].

4. How to include prosody?

4.1. Prosody in speech technology today

We have discussed the basic ideas of cognitive systems in detail as we believe that prosody (due to its communicative function) must be integrated in the system much more than it is in our recent speech technology products.

It must be stated that the most speech recognition systems do not use prosodic information explicitly. This is normally justified with the argument that this information is implicitly included in the feature vector of the recognizer. This is probably not generally true. E. g., we have demonstrated that the recognition of command words can be improved if their special prosodic emphasizing is considered [37].

If required, the recognition of emotional or other individual conditions is performed by means of separate, non-verbal recognizers. Their performance is continuously improving [38, 39], but the application is still limited to special cases (forensic applications, monitoring of automated telephone dialogues, etc.).

On the other hand, speech synthesis in practice is preferably text-to-speech (TTS) synthesis. Because a good prosody is crucial for the acceptance of synthetic speech, TTS was one of the driving forces for the rapid development of prosody research during the last two decades. The prosodic parameters of the synthesized speech signal (pitch, phone duration, sometimes also energy) are determined basing on the linguistic analysis of the input text and the available prosodic knowledge sources. However, coding prosody in written language is not straightforward. Therefore, finding the right prosody from text proves to be a very complex problem, which seems to be the reason why many systems use neural networks for this purpose.

There are many applications where an information retrieval system is equipped with speech output. In nearly all cases, the interface between both subsystems transfers simply text. Of course, this causes information losses compared to the integrated process (generation), but there are only few attempts to design information retrieval systems which are generating prosodic annotations [40].

An (at least potentially) full interface between speech recognizer and synthesizer on the one hand and some speech understanding system on the other hand will be normally found only in big research systems. The speech-to-speech translation prototype system Verbmobil [31] was designed to process even prosodic information in the speech understanding component [41].

The unsatisfactory integration of the prosodic processing in today’s systems is characterized in Fig. 3 by the additional components which “embrace” the core algorithms. This reflects the situation when UASR was invented but should not be applied to the improved design in Fig. 5. We could equip this flowchart with a similar “brace”, but with the purpose to separate irrelevant information. It depends on the task of the system, which information should be counted as irrelevant, and this is not automatically a prosodic one.

This means, that we have relevant parts of prosodic information which have to be integrated into the bottom-up process of speech analysis and must partially reach the speech understanding level. On the other side, if the speech understanding component generates some output, it should be enriched with everything what could be useful to simplify the top-down process where the right prosody is calculatied.

4.2. The role of analysis-by-synthesis experiments

The integration of prosodic information in both the bottom-up and the top-down branch is related to a number of research problems which we will mention here very briefly:

- The feature system which describes the prosodic effects at the different hierarchy levels is historically grown and rather complicated. It is optimized for the requirements of TTS systems. Therefore we can utilize many experiences from this field. However, partial inconsistencies and similar problems are expected if we want to fit the requirements of the structure in Fig. 5.
- Experience at the level of the feature vectors shows that it is not useful to combine the spectral and the prosodic features to a supervector. Multistream HMMs which are described in the literature [42] are known to be able to solve the problem. This approach proved also to be helpful in the recognition branch [43].
- The integration of prosodic features in the UASR structure leads to a growing complexity which will also influence the application of the FST technology. Some experiences on the application of FST for describing prosodic effects are available [44].
- A topic which we did not touch so far is the complex of speech rhythm. In TTS synthesis, the time structure of speech is considered by the duration of sounds and syllables. It is controversial discussed whether it is useful to include the analysis and synthesis of timing at higher levels in systems of speech technology, but the expectations are rather high, also due to biological arguments.

These (and other) questions can be answered only by means of experiments which need a platform which is algorithmically equipped for analysis-by-synthesis. Some demonstrations basing on UASR are available [45], but deeper investigations will need much more effort.

4.3. Projects which can support basic research

Although we can demonstrate numerous practical applications of the UASR technology during the last years, UASR and - even more - its “cognitive” successor are systems of basic research. Their development may benefit from other projects in speech technology which are more oriented to practical applications:

- Processing music signals is a rapidly growing field in acoustical signal processing. It is obvious that music and prosody research apply similar algorithms. In contrast to speech research, the importance of rhythm in music is not questionable at all. Therefore it will be useful to apply the encouraging developments from the musical rhythm analysis in the field of speech rhythm [46, 47].
- Teaching software for second language learners is another attractive field of software development which can also support prosody research. Our system AZAR was firstly applied for pronunciation training of German for users with Russian mother tongue [48] and later on for Chinese users [49]. Systems like AZAR improve the segmental quality of spoken language, but this can be extended to suprasegmental qualities [50]. Further users come from rehabilitation engineering (as for patients with Parkinson’s disease [51]) because their prosodic capabilities are limited. The development of these systems is connected to the acquisition of large prosodic databases, which can be used for basic research also.

5. Conclusions

We have shown that the epistemological approach of analysis-by-synthesis (AbS) did not lose the importance since the start of speech technology with the work of Kempelen. Prosody research has a very high complexity, and AbS is the proper method for optimizing its applications in speech and language engineering. We have shown that the AbS approach has largely contributed to the system theory, resulting in establishing the class of cognitive dynamic systems. Due to the hierarchical structure of speech and language, speech technology requires especially the introduction of *hierarchical* cognitive dynamic systems. Some aspects of this development have been reflected with respect to problems in prosody research.

6. Acknowledgements

The author wants to express his sincere thanks to the members of his team at the TU Dresden who developed UASR and many applications of it over a long time. This paper refers on their original papers according to the references. However, this paper is not only a summary of former work, but highlights also the “prosodic facet” of future plans. The author thanks therefore the members of the group who elaborated many ideas on future cognitive systems during the last two years: Matthias Wolff and Ronald Römer (now with TU Cottbus), Günther Wirsching (Katholische Universität Eichstätt), Robert Lorenz, Markus Huber, and Christian Kölbl (Universität Augsburg).

7. References

[1] Tscheschner, W., „Kommunikation und Kommunikationsgeräte”, Teaching Material, TU Dresden, 1982, 2nd ed. 1988.

[2] Hoffmann, R. and Mehnert, D., “Early experiments on prosody in synthetic speech“. In: Mixdorff, H. (ed.), Proc. 21th Conf. Electronic Speech Signal Processing, Berlin 2010, Studientexte zur Sprachkommunikation vol. 58, pp. 23 - 28.

[3] Kempelen, W. v., „Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine“. Wien: Degen, 1791.

[4] Mehnert, D. and Hoffmann, R., “Measuring pitch with historic phonetic devices”. In: Hoffmann, R. and Mixdorff, H. (eds.), 3rd International Conference on Speech Prosody, Dresden 2006, Studientexte zur Sprachkommunikation vol. 40, pp. 927 - 931.

[5] Schmidt, K.-O., „Verfahren zur besseren Ausnutzung des Übertragungsweges“, German Patent 594 976, patented February 27, 1932. - Supplementary Patent 722 607, patented January 14, 1939.

[6] Dudley, H. W., “Signaling System”, US Patent 2,098,956, patented Nov. 16, 1937.

[7] Hoffmann, R., “On the development of early vocoders”, Proc. of the 2nd IEEE Conference on the History of Telecommunications (Histelcon 2010) – A Century of Broadcasting, Madrid, 2010, pp. 359 – 364.

[8] Isačenko, A. V. and Schädlich, H.-J., „Untersuchungen über die deutsche Satzintonation“. Berlin: Akademie-Verlag 1964.

[9] Isačenko, A. V. and Schädlich, H.-J., “A Model of Standard German Intonation”. The Hague / Paris: Mouton 1970.

[10] Hoffmann, R., „Sprachsynthese an der TU Dresden: Wurzeln und Entwicklung“. In: Wolff, D. (ed.), Beitr. zur Geschichte u. neueren Entwicklung der Sprachakustik und Informationsverarbeitung, 2005, Studientexte zur Sprachkommunikation vol. 35,. pp. 55 - 77.

[11] Mehnert, D., „Grundfrequenzanalyse und –synthese der stimmhaften Anregungsfunktion, ein Beitrag zur Erzeugung und Verarbeitung sprachlicher Signale“. Dr.-Ing. thesis, TU Dresden 1975.

[12] Mehnert, D., „Analyse und Synthese suprasegmentaler Intonationsstrukturen des Deutschen, ein Beitrag zur Optimierung technischer Sprachkommunikationssysteme“. Habilitation thesis, TU Dresden 1985.

[13] Hamon, C., Moulines, E. and Charpentier, F., “A diphone synthesis system based on time-domain prosodic modifications of speech”. Proc. IEEE ICASSP, 1989, pp. 238-241.

[14] Fujisaki, H., “Information, prosody, and modeling - with emphasis on the tonal features of speech”. In: Fant, G., Fujisaki, H., Cao, J. and Xu, Y. (eds.): From traditional phonology to modern speech processing. Beijing 2004, pp. 111 - 128.

[15] Kruschke, H. and Koch, A., “Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization”, Proc. IEEE ICASSP, 2003, vol. 1, pp. 524 - 527.

[16] Hoffmann, R., Hirschfeld, D., Jokisch, O., Kordon, U., Mixdorff, H. and Mehnert, D., “Evaluation of a multilingual TTS system with respect to the prosodic quality”. Proc. of 14th International Congress of Phonetic Sciences, San Francisco, 1999, pp. 2307 - 2310.

[17] Holmes, J. N., “Speech Synthesis and Recognition”. London: Van Norstrand Reinhold 1988.

[18] Falaschi, A., Giustiniani, M. and Verola, M., “A hidden Markov model approach to speech synthesis“. Proc. Eurospeech, Paris, 1989, pp. 187 - 190.

[19] Tokuda, K., et al., “Speech parameter generation algorithms for HMM-based speech synthesis”. Proc. IEEE ICASSP, Istanbul, 2000, pp. 1315 - 1318.

[20] Eichner, M., Wolff, M. and Hoffmann, R., “A unified approach for speech synthesis and speech recognition using stochastic Markov graphs”. Proc. ICSLP, Beijing 2000, vol. 1, 701 – 704.

[21] Hoffmann, R., Eichner, M. and Wolff, M. “Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system”. In: Esposito, A., et al. (eds.): Verbal and Nonverbal Communication Behaviours. Berlin etc.: Springer 2007 (LNAI vol. 4775), pp. 200 - 218.

[22] Werner, S., Eichner, M., Wolff, M. and Hoffmann, R., “Towards spontaneous speech synthesis - Utilizing language model information in TTS”. IEEE Trans. on Speech and Audio Processing 12 (2004) 4, pp. 436 - 445.

[23] Wolff, M. and Hoffmann, R., “An Approach to Intelligent Signal Processing”. In: Esposito, A., et al. (eds.): Cognitive Behavioural Systems. Berlin etc.: Springer LNCS, 2012, 18 pp., in print.

[24] Strecha, G., Wolff, M., Duckhorn, F., Wittenberg, S. and Tschöpe, C., “The HMM synthesis algorithm of an embedded unified speech recognizer and synthesizer”. Proc. Interspeech, Brighton, 2009, pp. 1763 - 1766.

[25] Wolff, M. Kordon, U., Hussein, H., Eichner, M., Tschöpe, C. and Hoffmann, R., “Auscultatory blood pressure measurement using HMMs”. Proc. IEEE ICASSP, Honolulu 2007, vol. 1, pp. 405 - 408.

[26] Tschöpe, C. and Wolff, M., “Statistical classifiers for structural health monitoring”. IEEE Sensors Journal 9 (2009) 11, pp. 1567 - 1576.

[27] Haykin, S., “Foundations of cognitive dynamic systems”. IEEE Lecture, Queens University, 29 January 2009, http://soma.mcmaster.ca/papers/Slides_Haykin_Queens.pdf

[28] Haykin, S., “Cognitive radar”. IEEE Signal Processing Magazine 23 (2006) 1, pp. 30 - 40.

[29] Mitola III, J. and Maguire Jr., G. Q. “Cognitive radio: making software radios more personal”. IEEE Personal Communications Magazine 6 (1999) 4, pp. 13 - 18.

[30] Fuster, J. M., “Cortex and Mind - Unifying Cognition”. New York: Oxford University Press 2003.

[31] Wahlster, W., “Verbmobil – Foundations of Speech-to-Speech Translation”. Berlin etc.: Springer 2000.

[32] Wirsching, G., Huber, M., Kölbl, C., Lorenz, R. and Römer, R., "Semantic dialogue modeling“. In: Esposito, A., et al. (eds.): Cognitive Behavioural Systems. Berlin etc.: Springer LNCS, 2012, 11 pp., in print.

[33] Mohri, M. “Weighted automata algorithms”. In: Droste, M., Kuich, W. and Vogler, H. (eds.), Handbook of Weighted Automata. Monographs in Theoretical Computer Science. Berlin etc.: Springer, 2009, pp. 213 - 254.

[34] Wolff, M. „Akustische Mustererkennung“. Habilitation thesis, TU Dresden, 2011, Studientexte zur Sprachkommunikation, vol. 57

[35] Kölbl, C., Huber, M. and Wirsching, G., „Endliche gewichtete Transduktoren als semantischer Träger“. In: Kröger, B. and Birkholz, P. (eds.), Elektronische Sprachsignalverarbeitung 2011, Studientexte zur Sprachkommunikation vol. 61, pp. 176 - 183.

[36] Römer, R., “A cortical approach based on cascaded bidirectional hidden Markov models”. In: Esposito, A., et al. (eds.): Cognitive Behavioural Systems. Berlin etc.: Springer LNCS, 2012, 7 pp., in print.

[37] Kühne, M., Wolff, M., Eichner, M. and Hoffmann, R. “Voice activation using prosodic features”. Proc. Interspeech, Jeju, 2004, pp. 3001 - 3004.

[38] Schuller, B., Steidl, S. and Batliner, A., “The Interspeech 2009 Emotion Challenge”. Proc. Interspeech, Brighton, 2009, pp. 312 - 315.

[39] Schuller, B., et al., “The Interspeech 2010 Paralinguistic Challenge”. Proc. Interspeech, Makuhari, 2010, pp. 2794 - 2797.

[40] Schnell, M., „Prosodiegenerierung für die datenbasierte Sprachsynthese“. Dr.-Ing. thesis, TU Dresden, 2006, Studientexte zur Sprachkommunikation, vol. 38.

[41] Kompe, R., “Prosody in speech understanding systems“. Berlin etc.: Springer 1997 (LNAI vol. 1307).

[42] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., “Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis”. Proc. Eurospeech, Budapest, 1999, pp. 2347 – 2350.

[43] Janin, A., Ellis, D. and Morgan, N. “Multi-stream speech recognition: Ready for prime time?”, Proc. Eurospeech, Budapest, 1999, pp. 591 – 594.

[44] Hussein, H., Wolff, M., Jokisch, O., Duckhorn, F., Strecha, G. and Hoffmann, R., “A hybrid speech signal based algorithm for pitch marking using finite state machines“. Proc. Interspeech, Brisbane, 2008, pp. 135 - 138.

[45] Hussein, H., Strecha, G. and Hoffmann, R., “Resynthesis of prosodic information using the cepstrum vocoder“. Proc. 5th Int. Conf. on Speech Prosody, Chicago, 2010, paper 358 (4 pp.).

[46] Hübler, S. and Hoffmann, R., “Comparing the rhythmical characteristics of speech and music – Theoretical and practical issues”. In: Esposito, A., et al. (eds.), Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues. Berlin etc.: Springer 2011, pp. 376 – 386 (LNCS vol. 6456).

[47] Hübler, S. and Hoffmann, R. “A study on the metrical structure of music with similarity experiments”. Proc. 6th International Conference on Speech Prosody, Shanghai, 2012, this volume.

[48] Jäckel, R. and Hoffmann, R., „Sprachsignalanalyse für Menschen mit Deutsch als Zweitsprache. Eine intelligente Lerner-Software zur Aneignung der deutschen Standardaussprache. Deutsch als Zweitsprache 8 (2008) 3, pp. 37 – 47.

[49] Ding, H., Hoffmann, R. and Jokisch, O., “An investigation of tone perception and production in German learners of Mandarin”. Archives of Acoustics 36 (2011) 3, pp. 509 – 518.

[50] Jäckel, R. and Hussein, H., „Kontrastive Untersuchung zur Realisierung der Fokusakzente in gelesenen Äußerungen“. In: Hoffmann, R. (ed.), Elektronische Sprachsignalverarbeitung 2009, Studientexte zur Sprachkommunikation vol. 53, pp. 380 - 387.

[51] Ma, J. K.-Y. and Hoffmann, R., “Acoustic analysis of intonation in Parkinson’s disease”. Proc. Interspeech, Makuhari, 2010, pp. 2586 – 2589.