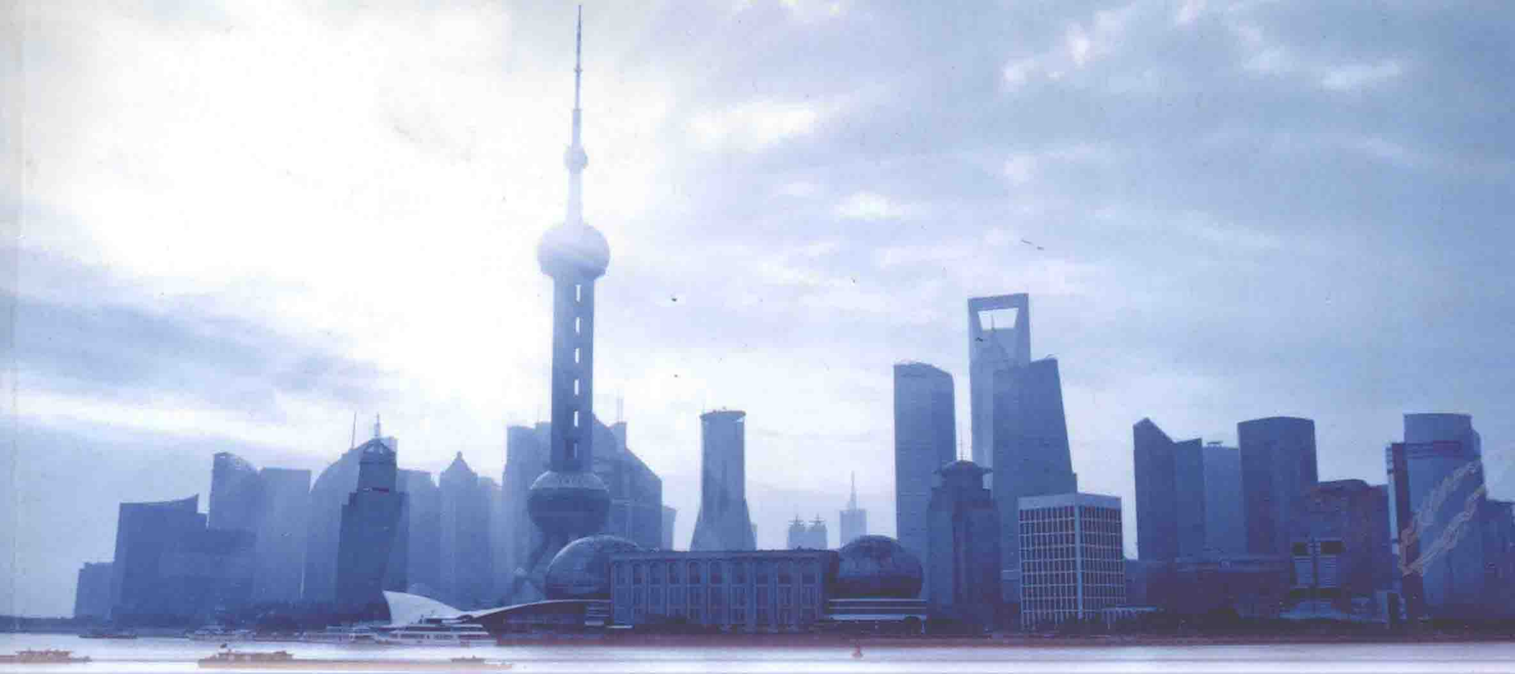

PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON SPEECH PROSODY

Shanghai, May 22-25, 2012

Qiuwu Ma, Hongwei Ding and Daniel Hirst (eds.)

Volume I

同濟大學出版社
TONGJI UNIVERSITY PRESS

PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON SPEECH PROSODY

Shanghai, May 22-25, 2012

Qiuwu Ma, Hongwei Ding and Daniel Hirst (eds.)

同濟大學出版社
TONGJI UNIVERSITY PRESS

图书在版编目(CIP)数据

第 6 届国际韵律大会论文集:Proceedings of the 6th International Conference on Speech Prosody 2012:英文/马秋武, 丁红卫,(法)赫丹(Hirst, D.)主编. --上海:同济大学出版社,2012.5

ISBN 978-7-5608-4869-3

Ⅰ.①第… Ⅱ.①马… ②丁… ③赫… Ⅲ.①韵律(语言)—国际学术会议—文集—英文 Ⅳ.①HO14-53

中国版本图书馆 CIP 数据核字(2012)第 088466 号

第 6 届国际韵律大会论文集

Proceedings of the 6th International Conference on Speech Prosody 2012

马秋武　丁红卫　Daniel Hirst　主编

Qiuwu Ma, Hongwei Ding and Daniel Hirst(eds.)

责任编辑 赵泽毓　**责任校对** 张德胜　**封面设计** 陈益平

出版发行　同济大学出版社　www.tongjipress.com.cn

(地址:上海市四平路 1239 号 邮编:200092 电话:021-65985622)

经　销　全国各地新华书店

印　刷　同济大学印刷厂

开　本　889 mm × 1240 mm　1/16

印　张　48

字　数　1 540 000

版　次　2012 年 5 月第 1 版　2012 年 5 月第 1 次印刷

书　号　ISBN 978-7-5608-4869-3

定　价　368.00 元

Table of Contents

Tuesday

PS2 Language Interaction and Acquisition; Prominence, Focus and Emphasis

PS2A Language Interaction and Acquisition

PS2B Prominence, Focus and Emphasis

PS4 Style, Attitude and Emotion; Syntax, Narrative, Speech Act and Discourse

PS4A Style, Attitude and Emotion

PS4B Syntax, Narrative, Speech Act and Discourse

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.

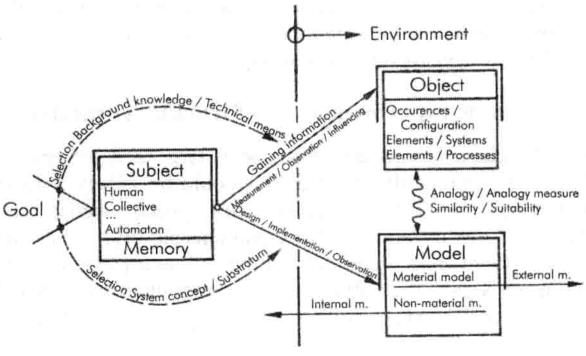

Figure 1: *Model method (adapted from* [1]*).*

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

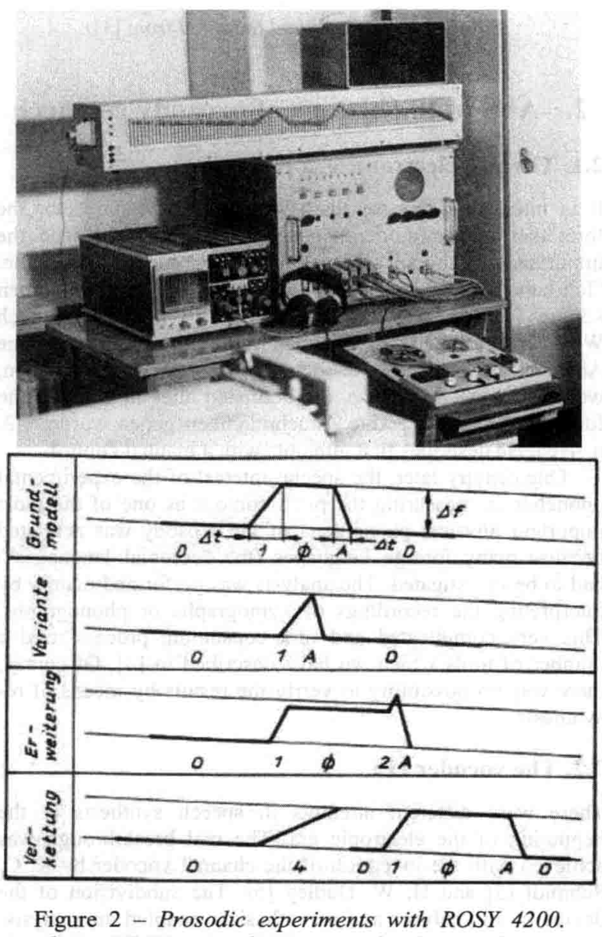

Figure 2 – *Prosodic experiments with ROSY 4200. Above: Experimental setup with the synthesizer terminal ROSY (middle right) and the contour generator (above). The control computer is not shown. – Below: Models of suprasegmental fundamental frequency contours from* [12].

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].

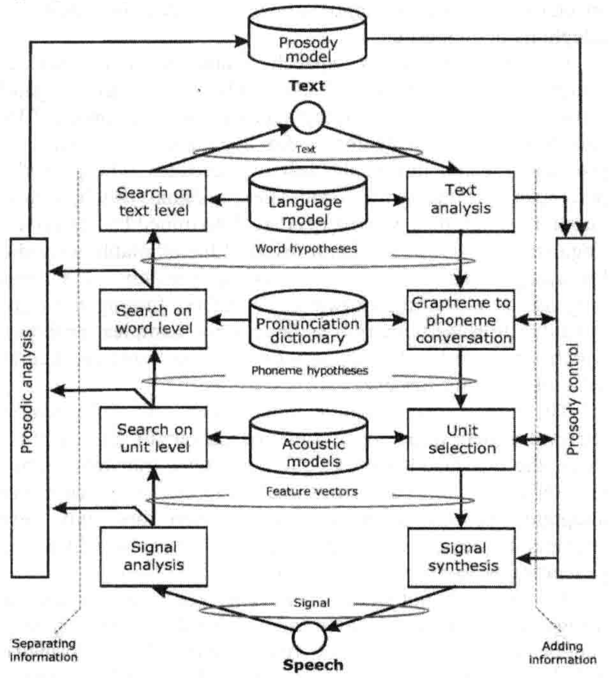

Figure 3: *Unified approach for speech recognition and synthesis (UASR), as it was proposed in the year 2000* [20].

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].

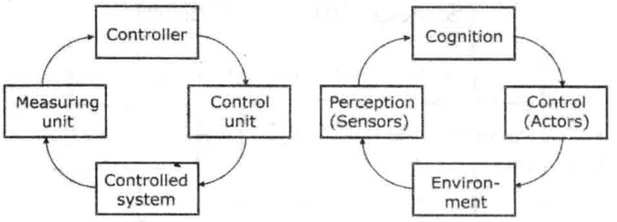

Figure 4: *Cognitive dynamic system as proposed by Haykin (right), compared with the loop of a classical system for automatic control (left).*

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.