EARLY EXPERIMENTS ON PROSODY IN SYNTHETIC SPEECH

Rüdiger Hoffmann and Dieter Mehnert

TU Dresden, Professur für Systemtheorie und Sprachtechnologie, D-01062 Dresden
ruediger.hoffmann@tu-dresden.de

Abstract: Synthetic speech needs prosody to get the right structure and to sound natural. Therefore, the emerging speech technology pushed the development of prosody models. Today, prosody research is well established with an own conference series, and powerful tools are available for investigating prosodic effects. The 80th birthday of the pioneer of quantitative prosody modeling, Professor Hiroya Fujisaki, is an excellent occasion to look at the situation in earlier times of speech technology. The authors give an outline using mainly the material which is available from the history in Dresden and Berlin. The oral presentation will be accompanied by numerous historic audio examples.

1 The pre-electronic era

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." [1, p. 413]. He describes first attempts with a manual control.

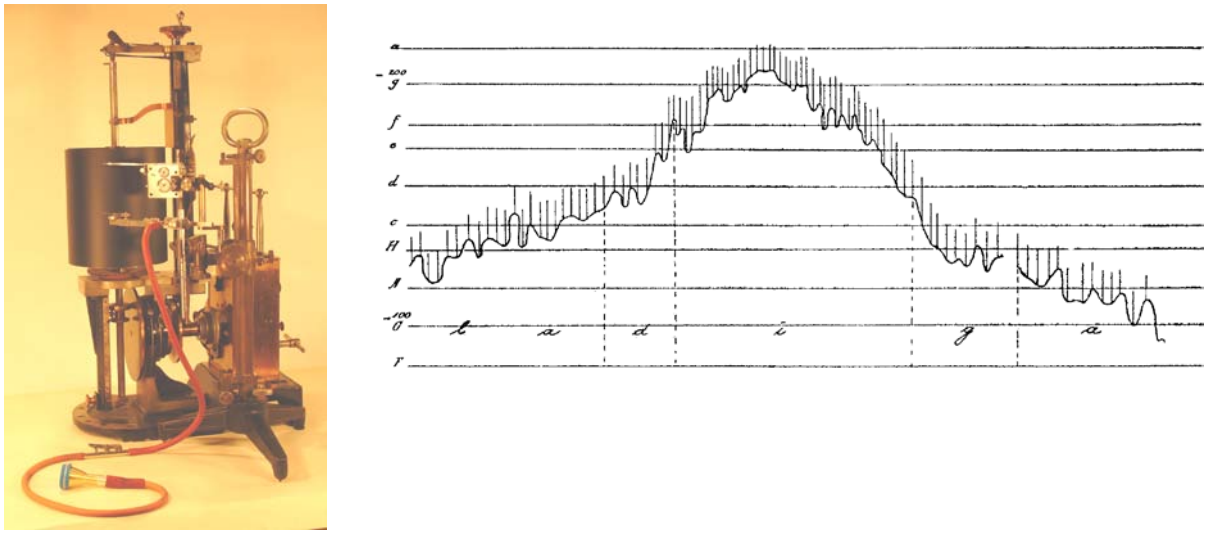

Figure 1 – Pre-electronic pitch recording. Left: Arrangement for recording the "throat sound" by a kymograph. Photograph from the historic acoustic-phonetic collection (HAPS) of the TU Dresden. – Right: Pitch contour ('la diga') produced by interpreting a kymographic recording [4].

100 years 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 (Figure 1). This very complicated and time-consuming process used a number of tools which we have described in [2, 3]. Of course, there was no possibility to verify the results by means of re-synthesis.

2 Analysis-by-synthesis: the vocoder

2.1 Development of the technology of channel vocoders

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 Dresden vocoder (Figure 2) which was developed by E. Krocker [8] are still available.

Figure 2 – Historic vocoders in Germany. Left – The Siemens vocoder in the background of the Siemens Studio for Electronic Musik, now in the Deutsches Museum in Munich. Right: Historic photograph of the Dresden vocoder. The left rack contained the analyzer channels, the two right racks the synthesizer channels [8].

2.2 The experiments from Isačenko and Schädlich

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 [9]. The English translation of their report [10] 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. (from [9]):

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.

3 Prosodic experiments with formant synthesizers

3.1 Development of formant synthesis

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. 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 [11]. The prosodic investigations which are described in the following section are 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.

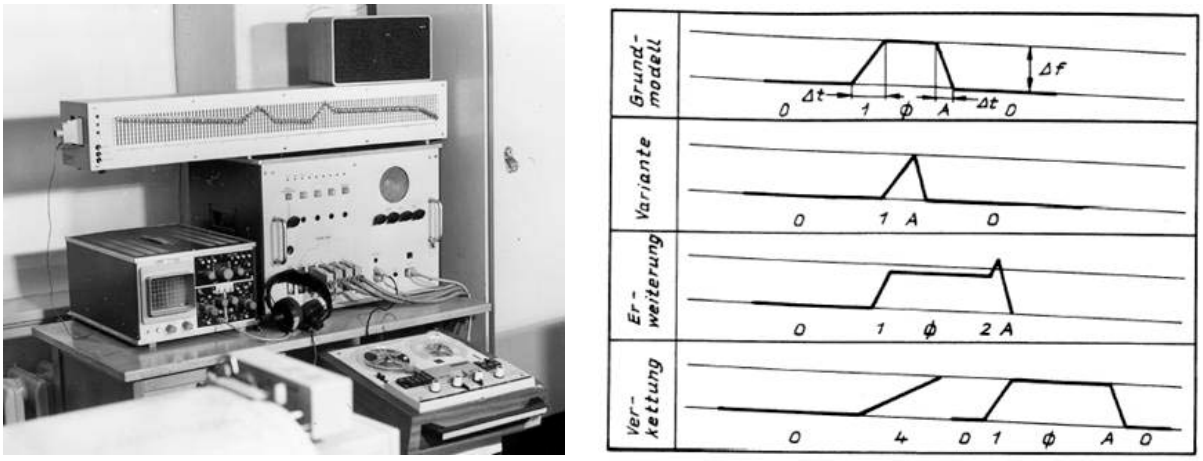

Figure 3 – Prosodic experiments with ROSY 4200. Left: Experimental setup with the synthesizer terminal ROSY (middle right) and the contour generator (above). The control computer is not shown. – Right: Models of suprasegmental fundamental frequency contours from [13].

3.2 Prosodic investigations

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 [12]. 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 [13] 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 3). Some examples of the test sentences in different intonation versions (monotonous / linear declination / declination plus accentuation) are still available as audio files.

4 Prosody in concatenative speech synthesis

4.1 Concatenation of waveforms in time domain

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. Single sounds which were naturally spoken could be stored and re-ordered into a new sequence. The synthesizer "Lora" is an early example. It consisted of a stapled series of storage elements like that in Figure 4. The different elements were equipped with pieces of magnetic tape storing the particular sounds. All elements were arranged in parallel, and the selection of the proper element was controlled in a complicated way using a camshaft. The main problem, however, was the production of naturally sounding sound transitions. It is reported that the transitions were implemented using the Schwa as intermediate sound [14].

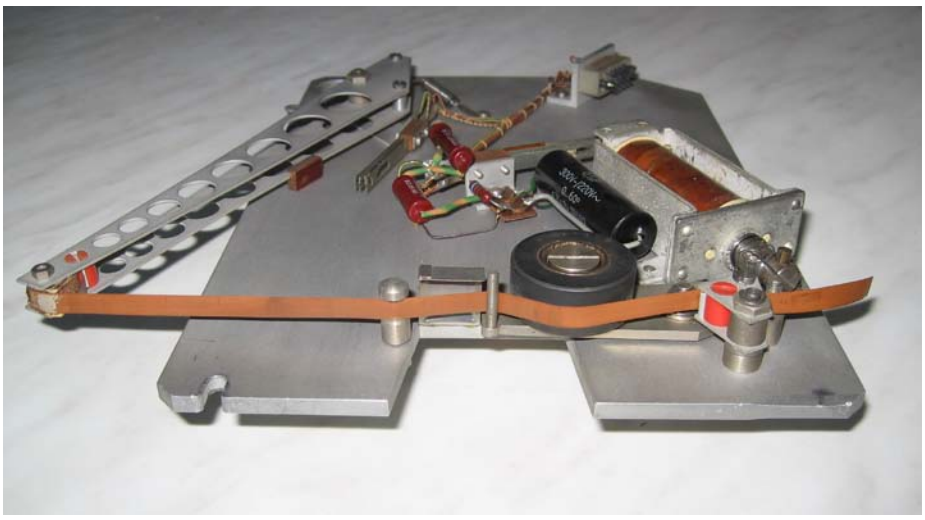

Figure 4 - A single segment of the magnetomechanical speech synthesizer "Lora" from 1964 (FTZ Darmstadt, Germany). One segment served as storage for one spoken sound. From the historic acoustic-phonetic collection of the TU Dresden.

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 [15] was the predominant solution und paved the way to a broad application of speech synthesis in time domain.

4.2 Prosody models for TTS systems, especially for DRESS

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 which was applied successfully to many languages. We describe briefly the adaptation of the model to the TTS system DRESS.

The Dresden Speech Synthesis System DRESS was developed in the 1990-th as a multilingual TTS system [16]. A number of algorithms for generating the pitch contour was compared [17], including the Fujisaki model which proved to be most applicable. Much effort was made to find effective training algorithms for the parameters of the Fujisaki model [18].

A more systematical investigation of the German prosody including the Fujisaki model was performed in the theses of H. Mixdorff [19, 20] with the development of the MFGI ("Mixdorff Fujisaki German Intonation") model. In this framework, we compared the prosodic quality of DRESS for different prosody models and found that MFGI performed favourable [21].

5 Conclusion

We have presented some historic examples of experimental prosodic investigations. The examples were selected with respect to the availability of historic audio material in the historic acoustic-phonetic collection (HAPS) of the TU Dresden.

The intonation model of H. Fujisaki had a strong influence at the development of the TTS technology in our university as well as worldwide. Recent activities in the Dresden group are continuously directed to the interplay of speech analysis and synthesis which includes the prosody. The application of the results is focused on embedded systems [22].

Bibliography

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

[2] Mehnert, D.; Hoffmann, R.: Measuring pitch with historic phonetic devices. In: Hoffmann, R.; Mixdorff, H. (Hrsg.): Speech Prosody. 3rd International Conference, Dresden, 2. - 5. 5. 2006. Dresden: TUDpress, 2006, pp. 927 - 931.

[3] Mehnert, D.: Prosodieforschung zu Beginn des 20. Jahrhunderts Tonhöhenmess- und Analysierverfahren mit historischen phonetischen Apparaten. In: Hentschel, C. (Hrsg.): Sprachsignalverarbeitung - Analyse und Anwendungen. Zum 65. Geburtstag von Klaus Fellbaum. Studientexte zur Sprachkommunikation, Bd. 44, pp. 28 - 45.

[4] Meyer, E. A.: Ein neues Verfahren zur grafischen Bestimmung des musikalischen Akzents. Monatszeitschrift für die gesamte Sprachheilpädagogik, 1922, p. 227-243.

[5] K.-O. Schmidt, 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. IEEE Histelcon, Madrid, 3. – 5. Nov. 2010, 6 p., in print.

[8] Krocker, E.: Aufbau und Untersuchung eines Übertragungssystems für synthetische Sprache, Dr.-Ing. thesis, TH Dresden, June 14, 1957.

[9] Isačenko, A. V.; Schädlich, H.-J.: Untersuchungen über die deutsche Satzintonation, Berlin: Akademie-Verlag 1964.

[10] Isačenko, A. V.; Schädlich, H.-J.: A Model of Standard German Intonation, The Hague / Paris: Mouton 1970.

[11] Hoffmann, R.: Sprachsynthese an der TU Dresden: Wurzeln und Entwicklung. Beitr. zur Geschichte u. neueren Entwicklung der Sprachakustik und Informationsverarbeitung, ed. by D. Wolf. Dresden: w.e.b. Universitätsverlag 2005, pp. 55-77 (Studientexte zur Sprachkommunikation, vol. 35).

[12] Mehnert, D.: Grundfrequenzanalyse und –synthese der stimmhaften Anregungsfunktion, ein Beitrag zur Erzeugung und Verarbeitung sprachlicher Signale. Diss. A, TU Dresden 1975.

[13] Mehnert, D.: Analyse und Synthese suprasegmentaler Intonationsstrukturen des Deutschen, ein Beitrag zur Optimierung technischer Sprachkommunikationssysteme. Diss. B (Habilitation thesis), TU Dresden 1985.

[14] Cramer, B.: Sprachsynthese zur Übertragung mit sehr geringer Kanalkapazität. Nachrichtentechnische Zeitschrift, vol. 17, 1964, pp. 413-424.

[15] Hamon, C.; Moulines, E.; Charpentier, F.: A diphone synthesis system based on time-domain prosodic modifications of speech. Proc. ICASSP, May 1989, pp. 238-241.

[16] Hirschfeld, D.: The Dresden Text-to-Speech System. In: Vích, R. (ed.): 6th Czech-German Workshop “Speech Processing”, Prague, 2.-4. 9. 1996. Prague: Academy of

Sciences, 1996, S. 22-24.
[17] Flach, G.; Kordon, U.: Generation of F0 contours for a German TTS system based on the Fujisaki model and neural nets. In: Vích, R. (ed.): 5th Czech-German Workshop “Speech Processing”, Prague, 27.-29. 9. 1995. Prague: Academy of Sciences 1995, pp. 26-27.
[18] Kruschke, H.; Koch, A.: Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), HongKong, China, 6.- 10. 4. 2003, vol. 1, pp. 524-527.
[19] Mixdorff, H.: Intonation patterns of German - Model-based quantitative analysis and synthesis of F0 contours, TU Dresden, Dr.-Ing. thesis, 26. 5. 1998.
[20] Mixdorff, H.: An integrated approach to modeling German prosody. Habilitation thesis. Dresden: w.e.b. Universitätsverlag, 2002 (Studientexte zur Sprachkommunikation, Band 25).
[21] Hoffmann, R.; Hirschfeld, D.; Jokisch, O.; Kordon, U.; Mixdorff, H. ; Mehnert, D.: Evaluation of a multilingual TTS system with respect to the prosodic quality. In: Proc. of 14th International Congress of Phonetic Sciences (ICPhS), San Francisco, 1.-7. 8. 1999, vol. 3, pp. 2307-2310.
[22] Hoffmann, R.: Speech synthesis on the way to embedded systems. In: Proc. of XI. International Conference Speech and Computer (SPECOM 2006), St. Petersburg, 25.-29. 6. 2006. keynote lecture. pp. 17-26.