TECHNOLOGY TRANSFER FOR SPEECH PROCESSING APPLICATIONS

Oliver Jokisch

Deutsche Telekom AG, Leipzig University of Telecommunication (HfTL)
Institute of Communications Engineering (ICE)
jokisch@hft-leipzig.de

Abstract: The article illustrates speech technology transfer of more than one decade. Some examples in speech synthesis, speech corpora development and pronunciation tutoring reflect experiences of GWT (the transfer partner of TU Dresden) and demonstrate a bidirectional know-how transfer between scientific and industrial partners.

1 Introduction

The contributions from Dresden in acoustics and speech analysis since 1911 – connected to professors like Heinrich Barkhausen, Walter Reichardt (from 1950) or Walter Tscheschner (from 1969) – are well-known and also referenced in this Festschrift. In speech processing, early developments involved prototypes in speech coding such as a vocoder in 1957 and several devices for parametric speech synthesis from 1962 to 1987 in cooperation with industrial partners (e.g. with GDR computer producer Robotron). This alliance of teaching, fundamental research and applied research has been continued by Prof. Rüdiger Hoffmann and his group since 1992. In cooperation with Fraunhofer institutes for applied research and corporate R&D departments, numerous projects in speech recognition, speech synthesis and database development were initiated. Additionally, Hoffmann supported the dissemination of speech technology by co-founding the companies Gesellschaft für Signalverarbeitung und Mustererkennung mbH (GeSIM) and voice INTER connect GmbH.

To ensure professional project management at the interface between university research and commercial application, he established the signal processing department within GWT-TUD GmbH (the technology transfer company of TU Dresden) in 1998. This article addresses technology transfer for speech processing applications and illustrates exemplary project results.

2 Knowledge and technology transfer in Dresden

2.1 Partnering concept

GWT-TUD GmbH – founded as knowledge and technology transfer company of TU Dresden – is acting as 'temporary company' for scientists of Saxonian universities. GWT realizes commercial projects and monitors time schedule, quality and financial budget. The service includes marketing, customer acquisition, contract negotiations, human and material resource management – while scientists can focus on research and development. GWT belongs to the leading enterprises of this kind in Germany. The service charge depends on project type (private, public or mixed funding), project complexity etc. and starts at ca. 12% of project’s gross total.
2.2 Exemplary research and development in speech processing

The following sections introduce three typical application areas in speech processing – speech synthesis, speech corpora development and pronunciation tutoring.

The described technology transfer initially based on fundamental and applied research in previous cooperation projects – such as the national Verbmobil project in the early 1990ies. The project results involved submodules e.g. text-to-speech (TTS) synthesis within a dialog system to demonstrate basic functions. Industrial project partners as e.g. the corporate research department of Daimler AG in Ulm (Germany) commanded a baseline TTS system in 1998 but specified additional requirements for real world applications such as a female corporate voice (for navigation system) with high intelligibility and naturalness or different interfaces and operation systems.

The specified tasks requested complex knowledge in research and development, an own experimental platform but also a lot of routine and optimization work which is not fully covered by academic research. Consequently, this knowledge and technology transfer was separately contracted by the industrial partners. In later projects with development departments of Infineon Technologies AG and Siemens AG (both located in Munich), we developed derivative TTS systems and submodules for the use in embedded environment under low-resource restrictions (memory on chip, calculation time etc.). The qualified feedback from the partners regarding the developed prototypes and applications was leading to new research questions at TU Dresden – in some cases concerning fundamental research as well. Additionally, prototypes were made available for teaching purpose so that the technology transfer as a whole can be considered as win-win situation for university research, commercial partners and students.

The quality of speech processing applications – in particular in speech synthesis – depends a lot of available speech corpora for training and localization/personalization. In a further section, the article shortly describes an according database project with Siemens CT in the multilingual environment. To demonstrate the cooperation with smaller companies in education and development, the next section illustrates some efforts in pronunciation tutoring technology which was developed in cooperation with the enterprises voice INTER connect GmbH and REZO Computerservice GmbH.

3 Speech synthesis technology

3.1 Dresden Speech Synthesizer (DRESS)

Stimulated by a speech synthesis solution on PCMCIA card (VoiceCard [1]) and the mentioned Verbmobil project (targeting on higher acoustic quality), several synthesis modules – including formant synthesis, concatenative (diphone-based) synthesis, text processing and prosody generation – have been developed [2].

The first GWT projects from 1998 till 2000 dealt with the software integration and improvement of these heterogeneous synthesis components for a general purpose text-to-speech (TTS) system. The design involved:

- Experimental setup (alternative modules and processing paths) and real-time capability,
- Multilingual databases and several voices,
- Strict code/data separation and different operating systems,
- Configurable modules for different application domains.
The resulting baseline system was called Dresden Speech Synthesizer (DRESS) [3, 4], based on concatenative PSOLA synthesis in time domain. Figure 1 shows the block diagram of DRESS. Main challenge was the interface adaptation for existing modules from different projects. Consequently, some modules formed bottlenecks with regard to processing quality and software technology – in particular on linguistic and prosodic level. During the multimedia and speech technology hype in the end of 1990s, several development projects in speech synthesis were acquired – including acoustic synthesis and voice optimization for Daimler AG, a TTS solution on Antares DSP board for the former T-Nova (Deutsche Telekom AG) and a TTS module for an interactive voice response (IVR) server of Globana Teleport GmbH. The hybrid multi-level prosody module in [3] has been patented by T-Nova.

3.2 Embedded speech synthesis – microDRESS and ePapageno

Around year 2000, TTS systems were considerably improved which resulted in memory requirements of several (partly a few hundred) megabytes – unacceptable for many applications on embedded systems. Such applications often require small footprint and low processing power (usually connected with restricted battery capacity). These requirements may be met if the text processing is changed from common data-driven algorithms to rule-based processing. Furthermore, the inventory should be as small as possible (diphone inventory) and should be stored in a compressed manner. In the mentioned cooperation with Infineon Technologies AG, a modified synthesis system microDRESS for German and English was developed [5] with a footprint of less than 1 megabyte per language. Simplified block diagram is shown in Figure 2.

The embedded target system of Infineon (see Figure 3) requested optimization steps like reduced computing power and small object code size, small external memory for compressed voice inventory and lexicons but also an optimized data flow and easy portability.
Compared to the baseline DRESS system, microDRESS does not show essential quality losses apart from the influences of telephone bandwidth which is appropriate for many embedded applications. Figure 4 compares the Mean Opinion Scores (MOS “overall quality”, scale from 1 “insufficient” to 5 “excellent”) for microDRESS stimuli with natural reference, DRESS stimuli and stimuli of two commercial TTS systems (marking typical quality expectation at this time). The test involved 20 listeners and 10 utterances. With uncompressed inventory, the quality of microDRESS (mdress) was not degraded compared to the baseline system DRESS (dress). An inventory coding similar to ADPCM (mdress*) caused an acceptable audible quality loss. The additional limitation to telephone bandwidth resulted in a clear quality loss of almost 1 MOS degree (mdress**). Indeed, experience showed that telephone bandwidth is widely accepted by users if presented via telephone set.

The modified TTS components (linguistic processing, prosodic model, unit inventory and inventory coding) were separately investigated but basing on a few assumptions [7], a cost-value metrics was predicted in terms of memory use versus auditory quality (Figure 5). It emphasizes that the memory resources for linguistic and prosodic processing are less critical. The unit inventory (including bandwidth and coding scheme) seems to be the cost-value bottleneck in scalable speech synthesis. Further Siemens projects were directed to a derived acoustic module for their corporate embedded synthesizer ePapageno – based on codecs from mobile communication. The involved one-step decoding and synthesis [8] was patented. The practical constraints of corporate R&D also stimulated the further research at TU Dresden. Advanced synthesis approaches of Strecha et al. achieve a memory consumption of less than 100 kB (code and data of complete acoustic synthesis).
4 Speech corpora and annotation

Multilingual localization and personalization in speech processing applications – in particular in speech synthesis – require the generation and evaluation of various speech databases – e. g. the voice inventory and a prosodic database for training. In cooperation with Siemens CT from 2002 till 2005, we prepared databases for several European languages.

First steps involved a casting of 5 to 8 female speakers per language. Figure 6 presents the perceptual ranking among six UK English speakers by preference decisions of 25 listeners. The correlation in decisions of native and non-native listeners (some not familiar with the target language) could be observed across all languages – except for a Mexican speaker in a Spanish casting (dialect effect) – and was surprisingly high. The intermediate evaluation results led to further research on objective (instrumentally measured) voice parameters correlating with listener’s preference as described in [9].

Figure 7 shows the averaged f0 mean values (pitch of the voice) of speakers with an equal rank 1 to 5 in German, UK and US English, Spanish and French as an example. The diagram suggests a cross-lingual preference for lower female voices.

Ensuring high quality of speech and descriptive data, advanced database projects require automatic procedures for collection, segmentation and annotation to reduce time effort and labor costs. Continuing the Siemens cooperation, we developed a semi-automatic software framework [10] involving manual segmentation and evaluation processes (overview in Figure 8) within the European project “Technology and Corpora for Speech to Speech Translation” (TC-STAR) in 2006, aiming at speech corpora for building advanced TTS systems as well as for intra-lingual/inter-lingual research on voice conversion and expressive speech. The resulting corpora for UK English, Spanish and Mandarin include voice recordings at 96 kHz/24 bit and consist of different read parts: e. g. novels and (expressive) parliamentary speech. In total, 10 hours of net speech per voice correspond to 90,000 running words in 5,558 utterances which needed to be annotated.

Phonetic and linguistic-prosodic annotation processes are denoted in the Figures 9 and 10. The different processing paths and interfaces enabled consistent manual and automatic labels. To support a robust prosodic segmentation, only minor and major phrases/accentes were marked.
5 Pronunciation tutoring technology

In the following, the technology transfer with local enterprises is illustrated. Further education company REZO Computerservice GmbH extended the professional training for Russian migrants with German language courses. The teachers were dissatisfied with existing computer-assisted tools which did not include a distinct user feedback. In cooperation with speech technology company voice INTER connect GmbH, we initiated two research projects 2004 – 2007 (funded by national agencies AiF and SAB) aiming at the "Automat for Accent Reduction" (German acronym: AzAR) [11]. The resulting software enabled the interactive training of second language (L2) pronunciation for native speakers from the Slavonic language group. Beside animated articulatory organs, position and strength of mispronunciation are represented by a color scale. Figure 12 shows the graphical user interface of AzAR. Despite a positive user feedback, the subsequent product marketing was less successful. Nevertheless, AzAR results led
to succeeding research projects under guidance of TU Dresden (e.g. Euronounce) and to new scientific partners, extended linguistic research and databases for Polish, Czech and Slovak.

**Figure 12:** User interface of AZAR 3, version 2010.

In a further step, didactic concept and tutoring technology were adapted to the special requirements of young children aged 3+ years. The prototypical embedded system LISA [12] – extended by microphone and loudspeaker – was integrated in a puppet and can lead simple dialogs. For navigation in educational books, a barcode scanner was connected (Figures 13 and 14).

**Figure 13:** Block diagram of LISA [12].

**Figure 14:** LISA demonstrator [12].

6 Conclusion

The examples of technology transfer demonstrate a fruitful and ‘bidirectional’ cooperation with commercial partners – often stimulating new fundamental research but also enhancing teaching horizon at TU Dresden. In the described cases, only pre-stages of products were achieved which requested further developments of the commercial partners.

7 Acknowledgment

I would like to thank my former colleagues at TU Dresden and GWT for their supportootnote{I joint the group of Prof. Rüdiger Hoffmann 1995 – 2012, worked as GWT project manager 1998 – 2007 and, at present, hold the Chair of Signal and System Theory at Leipzig University of Telecommunication (HfTL).} – in particular Prof. Rüdiger Hoffmann, PD Dr. Ulrich Kordon, Prof. Hongwei Ding, Dr. Diane Hirschfeld, Guntram Strecha, Dr. Sören Wittenberg, Rainer Jäckel and Henry Urban.
References


