Gaps to Bridge in Speech Technology

Géza Németh

Department of Telecommunications and Media Informatics (TMIT),
Budapest University of Technology and Economics, (BME) Hungary
nemeth@tmit.bme.hu

Abstract. Although recently there has been significant progress in the general usage and acceptance of speech technology in several developed countries there are still major gaps that prevent the majority of possible users from daily use of speech technology-based solutions. In this paper some of them are listed and some directions for bridging these gaps are proposed. Perhaps the most important gap is the "Black box" thinking of software developers. They suppose that inputting text into a text-to-speech (TTS) system will result in voice output that is relevant to the given context of the application. In case of automatic speech recognition (ASR) they wait for accurate text transcription (even punctuation). It is ignored that even humans are strongly influenced by a priori knowledge of the context, the communication partners, etc. For example by serially combining ASR + machine translation + TTS in a speech-to-speech translation system a male speaker at a slow speaking rate might be represented by a fast female voice at the other end. The science of semantic modelling is still in its infancy. In order to produce successful applications researchers of speech technology should find ways to build-in the a priori knowledge into the application environment, adapt their technologies and interfaces to the given scenario. This leads us to the gap between generic and domain specific solutions. For example intelligibility and speaking rate variability are the most important TTS evaluation factors for visually impaired users while human-like announcements at a standard rate and speaking style are required for railway station information systems. An increasing gap is being built between "large" languages/markets and "small" ones. Another gap is the one between closed and open application environments. For example there is hardly any mobile operating system that allows TTS output re-direction into a live telephone conversation. That is a basic need for rehabilitation applications of speech impaired people. Creating an open platform where "smaller" and "bigger" players of the field could equally plug-in their engines/solutions at proper quality assurance and with a fair share of income could help the situation. In the paper some examples are given about how our teams at BME TMIT try to bridge the gaps listed.

Keywords: Gaps in speech technology, domain-specific applications, open platform, user preferences.

1 Introduction

Speech technology has gained widespread use during my 30+ years in the area. From the appearance of modern personal computers there were exaggerating marketing

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 15–23, 2014.

predictions for exponential growth of speech technology (Fig. 1). This has never come true and although some people with vision such as Steve Jobs have seen the difficulties [2], it led to a roller-coaster type of investments and downgrading of speech R&D in the last three decades. There has been rather a linear increase of performance and acceptance of real-life applications in several countries worldwide.

	1981	1982	1983	1984	1985	AAGR (%) 1981-1985	1985 % OF TOTAL
SPEECH RECOGNITION							
Devices (Chips)	1	2	4	10	30	134%	20%
Products (Board Level)	10	17	36	70	100	78%	67%
Systems	4	6	9	13	20	50%	13%
Subtotal	$15	$25	$ 49	$ 93	$150	88%	100%
SPEECH SYNTHESIS							
Devices (Chips)	15	35	80	160	320	115%	65%
Products (Board Level)	5	12	25	50	100	111%	20%
Systems	3	9	20	40	75	124%	15%
Subtotal	$23	$56	$125	$250	$495	115%	100%
TOTAL	$38M	$81M	$174M	$343M	$645M	103%	

Source: Strategic, Inc.

Fig. 1. Speech technology market growth prediction between 1981-1985 [1]

Recently more realistic business predictions are presented [3] and some widely used applications are available in several countries (e.g. navigation systems, Apple's Siri, etc.). But there is still a long way to go in order to provide speech technology solution in most of the areas where human speech communication is used. Even in the most developed language and market (English) there are huge areas (e.g. language learning [4]) where the performance of current systems is not satisfactory. In this position paper I will introduce some of the gaps that I regard important to bridge in order to create systems that are more acceptable for the final judges, the end users.

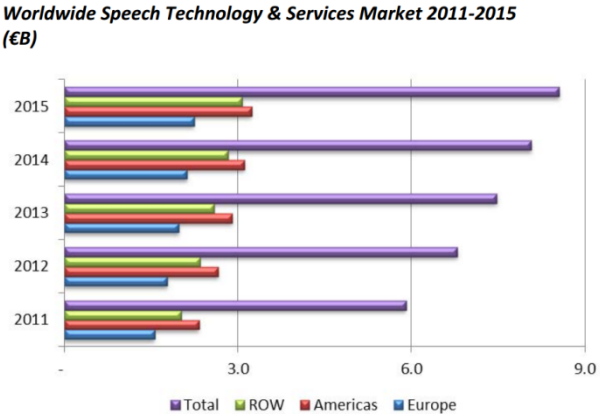

Fig. 2. Speech technology market growth prediction between 2011-2015 [3]

2 "Black-Box" Thinking

Perhaps the biggest gap to bridge for widespread use of speech technology is the education of both software developers/system integrators and end users. Both of them frequently consider TTS as a direct replacement of a "standard output device" (screen, printer or a human announcer) and ASR as a direct replacement for a "standard input device" (keyboard/mouse or a human typist). It is often ignored that typing errors are easy to detect when reading but when the mistyped text is read by a TTS system it may be very hard to comprehend. Similarly if the user mistypes something, a spell-checker may help. But a badly pronounced or out-of-vocabulary pronunciation cannot be corrected by the ASR module. There is an incredible amount of information that we use in human-human communication that is typically neglected in a speech technology application scenario. We know among others the age, education level, communication context, history of earlier communication, expertise, speaking rate of our partner and we can quickly adapt to all of these. So humans change both "their ASR and TTS" features significantly. Even during a single communication session we may request reading style change (e.g. ask for syllabification or spelling).

We have partially covered these needs in an e-mail reading application [5] by introducing three user levels (beginner, intermediate and expert) besides the chance to select the speaking rate. The verbosity of the menu system was adapted to the user level. Users also appreciated multiple system voices. In this e-mail reader application about 30% of the users changed the default male system prompt voice to a female alternative. In a reverse directory application [6] (input: phone number output: customer name and address is read out) the adaptation concept was implemented by three readout modes:

- continuous reading of the directory record (fast, overview mode)
- extended syllabification reading of the customer name (e.g. Bodó: Bo – Do with a long o) and continuous reading of the address (medium speed, supporting the detection of the exact written form of the customer name)
- spelling of the customer name character by character (slow, but very precise operation)

Users also prefer if the TTS system is not deterministic (i.e. not providing exactly the same waveform output for the same text input). We have found that such a solution can be implemented based on prosodic samples in various TTS technologies [7] with a definite user preference (c.f. Fig. 3). Our tests were performed for Hungarian and we are looking for interested partners to test the concept in other languages. It is important to note that speaking styles depend on voice timbre in addition to prosody, as well. So modelling voice timbre features (e.g. glottalization) is also an important topic [8].

Speech technology experts should be aware of and call the attention of the other contributing parties to these aspects and "educate" them about the optimal use of the available technology.

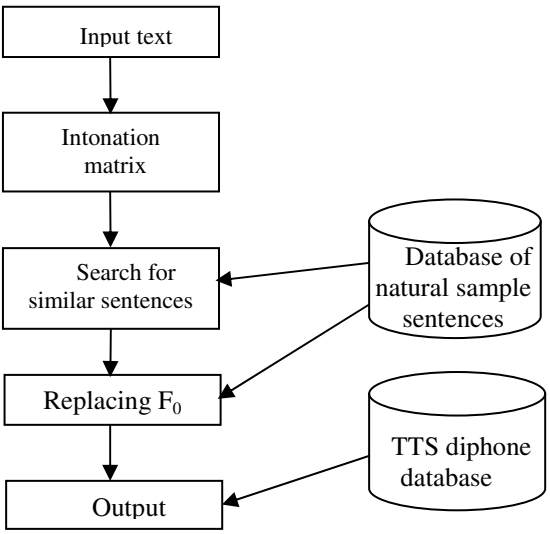

Fig. 3. The concept of generating TTS prosodic variability based on prosodic samples [7]

3 Generic vs. Domain Specific Solutions

Just as there is no single shoe type for everyone there is no single ASR or TTS system for all applications as long as we have no unified model of human communication suitable for engineering implementation. In the meantime the best approach is to create domain specific systems. It is not even sure that we should always strive for human-like performance. That may lead to the "uncanny valley" effect well known from robotics. Maybe in most cases our talking applications should behave rather in a way that resembles to special pets. This approach paves the way to the so-called etoinformatics.

Unfortunately for the time being there is not even a generic, standardized classification about the communicative contexts/speaking styles. ASR systems are still very much dependent on both the acoustic conditions and a priori knowledge about the recognition domain. More or less each research group develops its own alternatives. Associating the right application context to a particular technological solution may be critical from the end-users point of view. For example hyper-articulated script recordings may be optimal for intelligibility but may sound arrogant for the end-user. The most important factors for a given technology may also be domain/user dependent. For example to my great surprise several Hungarian blind users still prefer our 15 year-old diphone/triphone ProfiVox TTS system [9] as the Hungarian voice of the Jaws for Windows screen reader although there are other, newer Hungarian engines of international vendors. They have given the following justification:

- highly variable speech rate while maintaining good intelligibility
- fast response time (may be in the 10ms range)
- several voices (both male and female)
- optimized abbreviation handling.

The same system is also very well accepted as the voice of a humanoid robot [10]. But this system was completely unacceptable when presented as a mockup of a price-list reader over the telephone for a major telecom operator [11].

In the latter case the main requirement is that the output of the TTS system should not be easily distinguished from a recorded prompt and should be based on the voice talent of the company. A similar requirement applies to railway station announcements [12], and weather news reading [13]. In this case several hours of speech (in one of our applications more than 50 hours) has to be recorded and annotated in a corpus-based system in order to meet these requirements. This trend is expressed in the provisioning of several different system voices in the latest car navigation systems. Besides different voice timbre, dialects and various social speaking styles even with very harsh wording are provided as alternative speech output modalities. Recently in-car smart(phone) applications have gained a momentum after nearly 10 years of experimentation [14].

If the occasional clicks and glitches of corpus-based systems in case of out-of-domain text input is not acceptable or quick adaptation and creation of new voices is required than statistical parametric approach (usually HMM) is a trivial alternative. This solution can make use of already available tools and data created for waveform concatenation systems [15]. The output of the HMM system may be combined with higher quality elements of a corpus-based system so that this hybrid solution may only detected by expert evaluation. It is worthwhile to consider age related features as that may influence user preference as well [16]. TTS based expressive sound events –spemoticons- may offer a good link between objective and subjective aspects of sound perception [17].

4 “Large/Small” Languages

Of the 7106 known living languages of the world only 393 have more than one million first-language speakers [18]. There are only 23 languages with at least 50 million first-language speakers. According to the META-NET White Paper series on Europe’s Languages in the Digital Age [19] English is the only European language having good (not excellent) support in language and speech technology tools and resources. Central- and Eastern European languages mostly fall in the *fragmentary/weak/no support* with some *moderate* cases. During my Internet search I found less than 50 languages with TTS and less than 100 languages with ASR support worldwide. That does not include the domain specific alternatives that have been argued for in the previous sessions. So there is an incredible amount work that should be performed to provide proper solutions at least to several non-English speaking societies. There is a lack of readily available tools and data with specific information about language dependent and language independent features. Currently there is both lack of resources and multiplication of efforts to create the same (or similar) tools and resources.

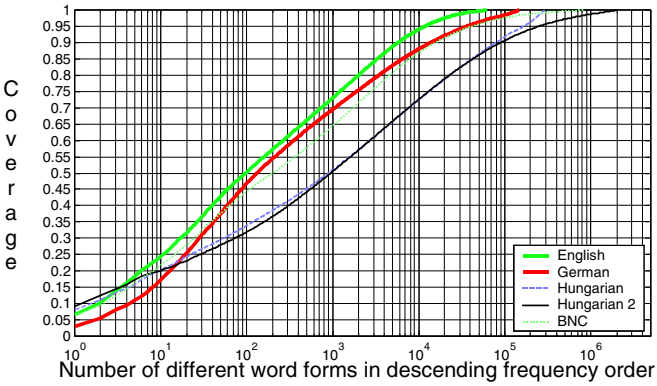

Fig. 4. Corpora coverage by the most frequent words (logarithmic horizontal scale) of standard texts [20]

A good illustration of this problem can be seen in Fig. 4. That study [20] investigated the number of different word forms (between space characters) appearing in various size of English, German and Hungarian corpora and the coverage that a given number of most frequent words can provide. It can be seen that the relatively small English corpus (3.5 million tokens) needs the least elements for a certain level of coverage. Hungarian has by far the largest number of words. That phenomenon trivially influences the vocabulary size for ASR systems but it is also exhibited in corpus-based TTS solutions. For example, in Hungarian more than 5.000 sentences were needed for proper coverage of the weather news domain which can be covered with about 1.000 sentences in English. A further problem for Hungarian is the relatively free word order.

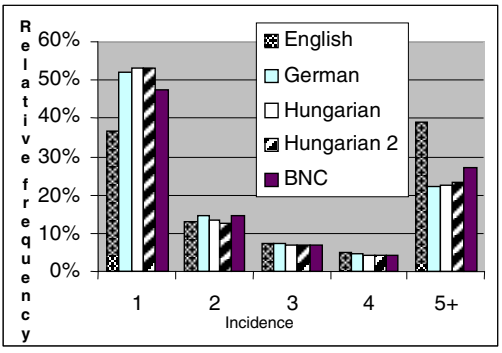

Fig. 5. Frequency of occurrences [20]

Figure 5 "illustrates a very problematic aspect –data scarcity- of corpus based approaches. It is clear, that even for English, which contained only 62.000 different word forms in a 3.5 million corpus, nearly 40% of the 62.000 different units (at least 20.000 words) appeared only once in the corpus. So even if one collects a huge corpus

for training a system, in case of a real-life application there is a very great probability that quite a few new items (related to the training corpus) will appear. If the corpus is large enough -such as the BNC for English- a very large ratio of rare items will appear only once. For Hungarian the problem is even harder. In a practically convincing case one should collect either such a big corpus, that all items should fall in the rightmost column (i.e. appearing at least five times in the corpus) or apply rule-based or other (non-word-based) approaches. Often the combination of several techniques may provide the best solutions."

The situation is very similar for Slavic languages, as well.

5 Closed and Open Platforms

It can be seen from the previous sections that no single company has the chance to cover at least the over 1million first-language speaker languages. Not to mention the several domain-specific adaptations, required for successful real-life applications. The only chance would be to use open platforms which allow the inclusion of new languages and domains for all respective application areas and provide quality assurance. Unfortunately currently there is a trend that operating systems manufacturers and systems integrators want to solve everything on their own or with some key partners. This approach prevents innovative academic groups, SME-s and non-profit civil organizations from developing new concepts that could be tested in real-life scenarios.

There is a great need for a really open environment where speech and language technology components can be efficiently integrated into new smart devices and environments. Experimenting with new solutions for disabled people is a critical factor because they may be highly dependent on an innovative engineering solution. For this reason they are ideal test subjects often with outstanding patience and thorough feedback. Current technological and policy limitations hinder advancement in several areas. The trend of the virtualization of services in infocommunication networks may open up new possibilities in this direction.

6 Conclusions

Although there has been enormous progress in speech technology during the last three decades there is still a long way to go. Due to the high dependence of speech communication on cognitive processes there is a very small probability of generic solutions for basic speech technology components in the foreseeable future. It seems to be more promising to create well-tailored engines for practically important applications. In order to be able to port solutions across platforms and languages it is important to define common scenarios (e.g. reading weather, scientific news, celebrity news, user manuals, directory services, subtitling of official vs casual conversations) and communicative contexts (e.g. informal-formal). For example both the health and the vehicle infotainment industries could be a good starting point. In an optimal case that could be implemented in an open testbed that would also provide quality assurance and testing. Due to a basically common cultural background with significant linguistic

and social variation Central- and Eastern Europe could be an optimal location for such an experiment. The BME TMIT team is open for co-operation from basic research to practical applications in any speech technology area.

Acknowledgments. The views expressed in the paper are those of the author. The research results mentioned and presented in the paper have been achieved by the co-operation of the Speech Communication and Smart Interactions Laboratory team at BME TMIT. They have been supported among others by the BelAmi, TÁMOP-4.2.1/B-09/1/KMR-2010-0002, CESAR (ICT PSP No 271022, EU_BONUS_12-1-2012-0005), PAELIFE (AAL_08-1-2011-0001), and the EITKIC_12-1-2012-0001 projects with support from the Hungarian Government, the Hungarian National Development Agency, the Hungarian National Research and Innovation Fund and the EIT ICT Labs Budapest Associate Partner Group.

References

1. Voice Synthesis Nearing Growth Explosion, Computerworld (August 31, 1981)
2. Brown, M.: The "Lost" Steve Jobs Speech from 1983; Foreshadowing Wireless Networking, the iPad, and the App Store. In: Talk by Steve Jobs at International Design Conference in 1983, October 2 (2012) (retrieved July 2014)
3. The Global Language Technology Market, LT-Innovate, p. 11 (October 2012)
4. Handley, Z.: Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Communication 51(10), 906–919 (2009)
5. Németh, G., Zainkó, C., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kiss, P.: The design, implementation and operation of a Hungarian e-mail reader. International Journal of Speech Technology 3/4, 216–228 (2000)
6. Németh, G., Zainkó, C., Kiss, G., Olaszy, G., Fekete, L., Tóth, D.: Replacing a Human Agent by an Automatic Reverse Directory Service. In: Magyar, G., Knapp, G., Wojtkowski, W., Wojtkowski, G., Zupancic, J. (szerk.) Advances in Information Systems Development: New Methods and Practice for the Networked Society, pp. 321–328. Springer (2007)
7. Németh, G., Fék, M., Csapó, T.G.: Increasing Prosodic Variability of Text-To-Speech Synthesizers. In: Interspeech 2007, Antwerpen, Belgium, pp. 474–477 (2007)
8. Csapó, T.G., Németh, G.: Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation. IEEE Journal on Selected Topics In Signal Processing 8(2), 209–220 (2014)
9. Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Gordos, G.: PROFIVOX - A Hungarian Professional TTS System for Telecommunications Applications. International Journal of Speech Technology 3(3/4), 201–216 (2000)
10. Csala, E., Németh, G., Zainkó, C.: Application of the NAO humanoid robot in the treatment of marrow-transplanted children. In: Péter, B. (ed.) 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), Kosice, Slovakia, pp. 655–658 (2012)
11. Németh, G., Zainkó, Cs., Bartalis, M., Olaszy, G., Kiss, G.: Human Voice or Prompt Generation? Can They Co-Exist in an Application? In: Interspeech 2009: Speech and Intelligence, Brighton, UK, pp. 620–623 (2009)

12. Klabbers, E.A.M.: High-Quality Speech Output Generation through Advanced Phrase Concatenation. In: COST Telecom Workshop, Rhodes, Greece, pp. 85–88 (September 1997)
13. Nagy, A., Pesti, P., Németh, G., Bőhm, T.: Design issues of a corpus-based speech synthesizer. HÍRADÁSTECHNIKA LX:(6), 6–12 (2005)
14. Németh, G., Kiss, G., Tóth, B.: Cross Platform Solution of Communication and Voice/Graphical User Interface for Mobile Devices in Vehicles. In: Abut, H., Hansen, J.H.L., Takeda, K. (eds.) Advances for In-Vehicle and Mobile Systems: Challenges for International Standards, pp. 237–250. Springer (2005)
15. Tóth, B., Németh, G.: Hidden Markov Model Based Speech Synthesis System in Hungarian. Infocommunications Journal LXIII:(7), 30–34 (2008)
16. Zainkó, C., Tóth, B.P., Bartalis, M., Németh, G., Fegyó, T.: Some Aspects of Synthetic Elderly Voices in Ambient Assisted Living Systems. In: Burileanu, C., Teodorescu, H.-N., Rusu, C. (eds.) Proceedings of the 7th International Conference Speech Technology and Human-Computer Dialogu, Cluj-Napoca, Romania, pp. 185–189. IEEE, New York (2013)
17. Németh, G., Olaszy, G., Csapó, T.G.: Spemoticons: Text-To-Speech based emotional auditory cues"m. In: ICAD 2011, Budapest, Magyarország, pp. 1–7. Paper Keynote 3 (2011)
18. Ethnologue, SIL International (retrieved July 2014)
19. META-NET White Paper series on Europe's Languages in the Digital Age (2013), http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison (retrieved July 2014)
20. Németh, G., Zainkó, C.: Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation. Acta Linguistica Hungarica 49:(3-4), 385–405 (2002)