(WHAT IS) THE CONTRIBUTION OF PHONETICS TO CONTEMPORARY SPEECH SYNTHESIS RESEARCH (?)

Petra Wagner

Universität Bielefeld
petra.wagner@uni-bielefeld.de

Kurzfassung: Recent advances in speech technology have significantly reduced the necessity for traditional phonetic system components or phonetic expertise, e.g. rule-based prosody models. We therefore need to ask the question, whether and how phonetics ought to play a role in ongoing and future speech synthesis development. The answer can be derived directly from a global analysis of the weaknesses in state-of-the-art systems. Their quality limitations unveil those areas where phonetics can still provide important impulses for system improvement. In this paper, it is argued that even if phoneticians "traditional" tasks have become less influential, with the advent of more dynamically flexible and interactive systems new challenges have arisen and need solutions within the humanities.

1 Phonetics in State-of-Art Speech Synthesis

Traditionally, phoneticians and engineers worked hand-in-hand on developing speech synthesis applications[1]: A lot of explicit phonetic knowledge was needed to hand-craft or statistically motivate rules in order to predict formant trajectories, phone durations, intonation contours and lots of other phonetic detail (e.g. [11] or the collection of many phonetically inspired approaches to synthesis in [27]). With data storage becoming less of a problem, and with data driven methods showing their superior ability to detect regularities and interdependencies in masses of data, phonetic knowledge was still needed when doing symbolic annotations or developing cost functions. In unit selection approaches, phoneticians determined the optimal levels of granularity for segmentation, prediction, unit selection or concatenation. In some sense, the phonetician's task has moved from *building* models him- or herself towards supervising the training of model building algorithms. But even the necessity for this type of expertise has been called into question:

> Fine grained phonetic data/topological rules are not necessarily required for generating plausible intelligible speech – as long as the carriers of those information are present in the text data (and corresponding speech recordings) used for training the system.' [25]

Thus, one could argue that whenever the need of building working applications rather than doing linguistic research is in the foreground, the phonetician's role has become minuscule at best, while the linguist remains somewhat needed for tagging the database. Roux's diagnosis seems to be correct when inspecting at the decreasing percentages of genuinely phonetic or linguistic publications at the latest international Speech Synthesis Workshops (cf. Table 1). It

[1] Many of the thoughts in this paper were presented as an invited talk at the ESSV 2011 meeting in Aachen and published as an abstract in [12]

needs to be stressed that the vast number of papers counted as "phonetic" are by no means non-technological. Rather, they were counted as "phonetic" because they all apply some phonetic insight in order to model the synthetic speech signal.

Table 1 - Relative and total numbers (in brackets) of phonetic or linguistic vs. purely technolgy-related contributions at international conferences focusing on aspects of TTS between 2004 and 2010.

Topic	SSW 5 (2004)	SSW 6 (2007)	SSW 7 (2010)
technology	30% (12)	38% (26)	64% (36)
phonetics/linguistics	70% (28)	62% (43)	36% (21)

The main reasons for this development lie (i) in state-of-the-art architecture specific challenges and chances, (ii) in scientific progress, and (iii) in a shift of interests within those parts of phonetics community sharing an interest in technical systems:

1. Naturally, the main scientific visions and goals of unit selection approaches lie in their data-orientation lending to the idea of using easily available resources such as audio books [4, 24], building on the natural phonetic variety available in large speech corpora (e.g. [6]), or using available resources for building applications for minority languages ([26, 18]). A main problem within statistical-parametric (HMM-based) approaches to speech synthesis [3] is the need to improve the voice quality. The architecture specific chances lie in the possibility to exploit the parametric approach by working on voice conversion algorithms and on modeling the expressive, emotional and gender specific characteristics of speech (e.g. [21, 20]). Despite the clear phonetic relevance of such topics, the solutions are at best marginally touching on phonetic aspects.

2. Progress in automatic labeling has made much of the former manual annotation work superfluous, especially with the advent of semi- or unsupervised machine learning techniques that need only few amounts of training data (e.g. [33]). Furthermore, much progress has been made within objective evaluation procedures, making speech synthesis research more independent of time consuming (phonetic) evaluation work (e.g. [19]). Also, synthesis systems have reached a very good quality, in some applications even outperforming humans, e.g. when being used for announcements in noisy environments such as train stations, where a somewhat machine-like hyperarticulation can be advantageous.

3. Probably linked to the good quality of many TTS systems, the research focus within the speech technologically interested phonetics community has shifted towards other applied research questions, such as using phonetic knowledge in CALL applications (e.g. [10]), modeling human speech processing using embodied, articulatory and neuronal approaches (e.g. [13, 28]), modeling the prosody of discourse and dialogue (e.g. [30, 7]), classifying phenomena of turn organization, overlap and pausing (e.g. [9, 34]), investigating communicative functions of non-verbal expressions such as laughter (e.g. [31]), modeling processes of communicative adaptation such as interactional convergence or entrainment (e.g. [14, 32, 15]) or multimodality (e.g. [29, 2]). Thus, building on a high-quality basic systems, researchers are currently aiming at making the technological applications more communicative, interactive, adaptive, expressive and multimodal.

In sum, it can be said that while it is true that the current fundamental research in speech synthesis are being made largely without phonetic and linguistic competence, the unsolved problems

such as building adaptive and interactive systems are still in need of a certain amount of insight probably gained from the humanities. The good news is that it has been the technological progress that has given phoneticians the freedom to tackle these issues. In the following section, I will discuss the current state of the art of speech synthesis applications and make a suggestion on how a reintegration of the technologically inspired communicative trend in phonetics can help improving current synthesis applications.

2 A Qualitative Assessment of State-of-the-Art Synthesis Systems

Speech synthesis has reached a plateau of quality, perhaps now being in a situation similar to speech recognition. In certain fields, synthesis has reached a high degree of acceptance, e.g. in announcement systems at train stations, in car navigation systems, computer games or mobile phone applications. Thus, we have empirical proof that users are able to adapt to synthetic speech even if it is still not perfect in the sense of "human". Interestingly, we do not really know why this is the case and how the systems could be further improved. Despite some success stories of working applications, many of our ongoing research concentrates on the boundaries of the technologically possible without knowing where this, e.g. highly emotional, authentic speech, would be useful in an application. Clearly, research should not be entirely application driven. Still, in a research field focusing on an application, it comes as a surprise that the question of applicability is often ignored. To this day, we have little or no knowledge about the individual user's needs and expectations:

> Even if a model can predict user judgements of "overall quality" with high validity and reliability, this is not necessarily a good indicator of the acceptability of a service. [...] Different users may focus on different quality dimensions in different contexts and weight them according to the task, context of use, price etc. [17]

Indeed, we need to get a understanding of our "typology of users", instead of tuning our applications and evaluations to the "average user" – as it is done by MOS approaches to system evaluation. In some sense or other, most users will belong to a "special user group", e.g. by being young or senior citizens, by being male or female, by being upset, bored, distracted or attentive. In fact, any potential user is likely to approach a system with "peculiar" or somewhat "special" personality or context dependent preferences or dislikes. Just as different people use different sets of options in their operating systems or mobile phones, they might want to use a different set of options in their speech synthesis systems. It is the task of researchers to find out what a good set of options and settings might be.

Some first attempts are being made to better grasp the user needs and expectations, e.g. by investigating preferences within or across special user groups such as blind users or senior citizens (e.g. [16, 35]). Besides the individual differences, typical contexts influencing our system demands need to be better understood. Such contexts could be situations where the user is distracted, e.g. while driving, or highly attentive, e.g. while interacting with a computer game, but there are only few studies of human-human listening behaviour in distracted vs attentive situations (cf. [5] and references therein). A very promising and rather active line of research lies in the better understanding of felicitous system feedback, e.g. using non-verbal expressions such as laughter, verbal or visual backchannels signalling a degree of system attention, similar to a human listener (e.g. [23, 8]). A better understanding of feedback may enable a more efficient communication between human and machine [1]. A better understanding about the adaptations between interlocutors, but also a better understanding of how recipient design, i.e. the way we adapt ourselves to the needs of our interlocutors, may be very helpful to improve these dynamical aspects in human-machine interaction in the future.

3 Discussion

A key problem in improving our current synthetic speech is our lack of understanding our user's needs. Certainly, many insights can be gained from studying human-human interaction but possibly, that approach leads us to wrong conclusions as humans probably do not expect human-like behavior from a machine. While WOZ-scenarios may tell us something about the potential interaction patterns we need to integrate into our systems, these share a problem with many of the recordings made to study human-human interaction (cf. Figure 1): They are carried out in laboratory contexts that are likely to elicit behaviors differing strongly from those in everyday situations. It would be advantageous if users could be studied using their synthetic speech devices in everyday applications. The most problematic aspect of less experimental control is a lack of finding systematic variables influencing the users' behavior. Therefore, just as we need to find better ways of studying authentic human-human-behaviour, we need to find better, systematic ways of evaluating our applications embedded in authentic situations thus making our findings ecologically more valid. Besides investigating better techniques of systematically capturing authentic user-system interactions in a minimally invasive manner, we need to better exploit existing technologies to deal with noisy data – as are typical for authentic situations. Progress in channel separation, audio signal enhancement need to be taken into account if we want to move our system evaluations "out of the lab and into the wild". Besides the technical issues, more work needs to be done to build annotation tools and standards for "noisy speech" occurring out of the lab. Based on these data, a kind of "user typology" may be developed which helps us to determine the system parameters that need to be taken into account when building working applications for "special user groups". The MOS as the still most widely used tool to investigate the quality of synthetic speech, does not appear to be the optimal method for system evaluations, unless we are paying close attention to the "outliers". The expertise of phoneticians should be very helpful in developing alternative methods for system architecture evaluation and in detecting the potential keys to a system's success or failure.

Figure 1 - The Figure on left shows a corpus recording environment where it has been tried to capture video, audio and motion data in a maximally authentic, minimally invasive manner [22]. The picture on the right shows a typical corpus recording under laboratory conditions [5].

4 Conclusion

This article built on the observation that phonetics has become less influential in building synthetic speech synthesis applications, as the state-of-the-art technologies used are less in need of phonetic models predicting the detailed sound structure of the synthesized units. However,

I have argued that with the increased quality of available system's, other, new aspects of system quality need to be addressed in speech synthesis development, which lie more in the area of adequateness in the human-system interaction. While engineers can build and provide the platforms on which appropriate technical systems are built, the way they are of use in human-machine interaction needs insight gained from the humanities. Those can concentrate on the following questions:

- What contexts (speaker or situation specific) exist that need an adjustment of the synthetic voice and style?
- What are the phonetic system parameters that should be subject to a context specific adjustment?
- What kind of phonetic adjustments are needed and useful?

References

[1] BAUMANN, T. and D. SCHLANGEN: *INPROiSS – A component for just-in-time incremental speech synthesis*. In *ACL*, pp. 103–108, 2012.

[2] BERGMANN, K., V. AKSU and S. KOPP: *The relation of speech and gestures: temporal synchrony follows semantic synchrony*. In MALISZ, Z., C. KIRCHHOFF and P. WAGNER (eds.): *Proceedings of GESPIN*, http://gespin.uni-bielefeld.de/?q=node/66, 2011.

[3] BLACK, A. W., H. ZEN and K. TOKUDA: *Statistical parametric speech synthesis*. In *Proceedings of ICASSP*, pp. 1229–1232, 2007.

[4] BREUER, S., S. BERGMANN, R. DRAGON and S. MÖLLER: *Set-up of a Unit Selection Synthesis with a Prominent Voice*. In *Proceedings of LREC 2006*, http://www.lrec-conf.org/proceedings/lrec2006/, 2006.

[5] BUSCHMEIER, H., Z. MALISZ, M. WŁODARCZAK, S. KOPP and P. WAGNER: *'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness*. In *Proceedings of Interspeech*, pp. 2057–2060, Florence, Italy, 2011.

[6] CAMPBELL, N.: *Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech*. IEICE Trans. Inf. Syst., E88-D(3):376–383, 2005.

[7] CHARFUELAN, M., M. SCHRÖDER and I. STEINER: *Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings*. In *Proceedings of Interspeech*, pp. 2558–2561, 2010.

[8] GRANSTRÖM, B. and D. HOUSE: *Inside out – Acoustic and visual aspects of verbal and non-verbal communication*. In *Proceedings of the 16th Congress of the Phonetic Sciences*, pp. 11–28, 2007.

[9] HELDNER, M. and J. EDLUND: *Pauses, gaps and overlaps in conversations*. Journal of Phonetics, 38:555–568, 2010.

[10] JOKISCH, O., A. WAGNER, R. SABO, R. JAECKEL, N. CYLWIK, M. RUSKO and R. HOFFMANN: *Multilingual Speech Data Collection for the Assessment of Pronunciation and Prosody Training in a Language Learning System*. In *Proceedings of Speech and Computer (SPECOM)*, pp. 515–520, 2009.

[11] KLATT, D.: *Synthesis by rule of Segmental Durations in English Sencences*. In LINDBLOM, B. and S. ÖHMANN (eds.): *Frontiers of Speech Communication Research*, pp. 287–299. London: Academic Press, 1979.

[12] KRÖGER, B. and P. BIRKHOLZ (eds.): *Elektronische Sprachsignalverarbeitung 2011 – Tagungsband der 22. Konferenz*, vol. 61, Studientexte zur Sprachkommunikation, Aachen, September 2011. Dresden, TUD Press.

[13] KRÖGER, B., J. KANNAMPUZHA and C. NEUSCHAEFER-RUBE: *Towards a neurocomputational model of speech production and perception*. Speech Communication, 51:793–809, 2009.

[14] LEVITAN, R. and J. HIRSCHBERG: *Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Dimensions*. In *Proceedings of Interspeech*, pp. 3081–3084, Florence, Italy, 2011.

[15] LEWANDOWSKI, N.: *Talent in nonnative phonetic convergence*. PhD thesis, Universität Stuttgart, URN: urn:nbn:de:bsz:93-opus-74023, 2012.

[16] MOERS, D. and P. WAGNER: *Assessing the adequate treatment of fast speech in unit selection systems for the visually impaired*. In *Proceedings of SSW6*, pp. 282–287, 2007.

[17] MÖLLER, S., K.-P. ENGELBRECHT and R. SCHLEICHER: *Predicting the Quality and Usability of Spoken Dialogue Systems*. Speech Communication, 50:730–744, 2008.

[18] NIEKERK, D. R. VAN and E. BARNARD: *Phonetic alignment for speech synthesis in under-resourced languages*. In *Proceedings of Interspeech*, pp. 880–883, 2009.

[19] NORRENBROCK, C., F. HINTERLEITNER, U. HEUTE and S. MÖLLER: *Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals*. IEEE Signal Proc. Letters, 19(5):255–258, 2012.

[20] NOSE, T. and T. KOBAYASHI: *Recent development of HMM-based expressive speech synthesis and its applications*. In *Proceedings of APSIPA ASC*, p. 189, 2011.

[21] NOSE, T. and T. KOBAYASHI: *Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency*. Speech Communication, 53(7):973–985, 2011.

[22] OERTEL, C., F. CUMMINS, J. EDLUND, P. WAGNER and N. CAMPBELL: *D64 – A corpus of richly recorded conversational interaction*. Journal of Multimodal User Interfaces, pp. 1–10, 2012.

[23] POPPE, R., K. TRUONG, D. REIDSMA and D. HEYLEN: *Backchannel Strategies for Artificial Listeners*. Intelligent Virtual Agents, pp. 146–158, 2010.

[24] PRAHALLAD, K. and A. W. BLACK: *Segmentations of monologues in audio books for building synthetic voices*. IEEE Trans. Audio, Speeech and Language Processing, 19(5):1444–1449, 2011.

[25] ROUX, J. C.: *Do we need linguistic knowledge for speech technology applications in African languages?*. In PAUW, G. DE, H. GROENEWALD and G.-M. DE SCHRYVER (eds.): *AFLAT 2010 – Proceedings of the Second Workshop on African Language Technology*, p. 1, http://aflat.org/aflat2010, 2010. European Language Resources Association (ELRA).

[26] ROUX, J. C. and A. S. VISAGIE: *Data-driven approach to rapid prototyping Xhosa speech synthesis*. In *Proceedings of SSW6*, pp. 143–147, 2007.

[27] SANTEN, J. P. VAN, R. W. SPROAT, J. P. OLIVE and J. HIRSCHBERG (eds.): *Progress in Speech Synthesis*. Springer: New York, 1997.

[28] ŠIMKO, J. and F. CUMMINS: *Sequencing and Optimization within an Embodied Task Dynamic Model*. Cognitive Science, 35(3):527–562, 2011.

[29] SWERTS, M. and E. KRAHMER: *Facial expressions and prosodic prominence: Comparing modalities and facial areas*. Journal of Phonetics, 36(2):219–238, 2008.

[30] SYRDAL, A., A. CONKIE, Y.-J. KIM and M. BEUTNAGEL: *Speech acts and dialogue TTS*. In *Proceedings of SSW7*, Kyoto, Japan, 2011.

[31] TROUVAIN, J. and N. CAMPBELL (eds.): *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, http://www.coli.uni-saarland.de/conf/laughter-07/proceedings.html, 2007. Saarland University, Germany.

[32] WAGNER, P., B. INDEN, Z. MALISZ and I. WACHSMUTH: *Interaction Phonology*. In WACHSMUTH, I., P. JAECKS and J. DE RUITER (eds.): *Towards a New Theory of Communication*. Benjamins, to appear.

[33] WATTS, O.: *Unsupervised Learning for Text-to-Speech Synthesis*. PhD thesis, University of Edinburgh, http://www.cstr.ed.ac.uk/publications/, 2012.

[34] WŁODARCZAK, M., J. ŠIMKO and P. WAGNER: *Temporal entrainment in overlapped speech: A cross-linguistic study*. In *Proceedings of Interspeech*, 2012.

[35] WOLTERS, M., P. CAMPBELL, C. DEPLACIDO, A. LIDDELL and D. OWENS: *Making Synthetic Speech Accessible to Older People*. In *Proceedings of SSW6*, 2011.