The speech synthesis phoneticians need is both realistic and controllable

Zofia Malisz[1], Gustav Eje Henter[1], Cassia Valentini-Botinhao[2], Oliver Watts[2], Jonas Beskow[1], Joakim Gustafson[1]
[1]*Department of Speech Music and Hearing, KTH, Stockholm, Sweden*
[2]*The Centre for Speech Technology, The University of Edinburgh, UK*
[malisz, ghe, beskow, jkgu]@kth.se, [cvbotinh, owatts]@inf.ed.ac.uk

Abstract

We discuss the circumstances that have led to a disjoint advancement of speech synthesis and phonetics in recent decades. The difficulties mainly rest on the pursuit of orthogonal goals by the two fields: realistic vs. controllable synthetic speech. We make a case for realising the promise of speech technologies in areas of speech sciences by developing control of neural speech synthesis and bringing the two areas into dialogue again.

Introduction

Text-to-speech (TTS) synthesis has made enormous progress recently. Modern synthetic voices sound nowhere close to the classic systems of yore – they have improved greatly in intelligibility and realism and are well on their way to achieve both expressivity and flexible speech style adaptation (Skerry-Ryan et al., 2018; Wang et al., 2018). This latest leap was ushered by powerful methods in machine learning, particularly deep learning, on large amounts of data (Watts et al., 2016).

Notably, however, the foundations of the progress lie in a close research and development loop between speech scientists and speech engineers that existed for decades (King, 2015). This has been particularly true for explicit acoustic and linguistic feature modelling and for evaluation standards. Unfortunately, at least since the advent of concatenative TTS (“cut-and-paste” methods) and certainly the machine learning revolution, the two fields have been growing apart. In this work, we concentrate on a particular trade-off that has arisen from the distance between the disciplines: realism vs. control.

It soon became evident in engineering that ever greater realism can be achieved at the expense of explicit modelling. The less explicit the modelling and the more data put into the ever more powerful machine learning algorithms, the better the performance of the synthesiser. Developers greatly diminishing focus on modelling has had important consequences for speech sciences, as it is the explicit modelling of particular acoustic and linguistic parameters that enables the creation or manipulation of synthetic output, i.e. control over the output speech.

Control over numerous meaningful, relatively low-level signal properties such as pitch, VOT, etc. has been invaluable for phonetic research in the past. Pertinent uses in speech sciences include experiments with synthetic stimuli in speech perception research. Important insights into phonetics, such as evidence for categorical speech perception, were reached with the use of synthetic sound continua (Lisker and Abramson, 1970). Theoretical advances such as the motor theory of speech perception (Lieberman and Mattingly, 1985) and acoustic cue analysis were also made possible by experiments with synthetic stimuli. Where empirical paradigms demand it, speech distortion and de-lexicalisation, or removal of cues to whole particular structures, such as prosody, are achieved us-

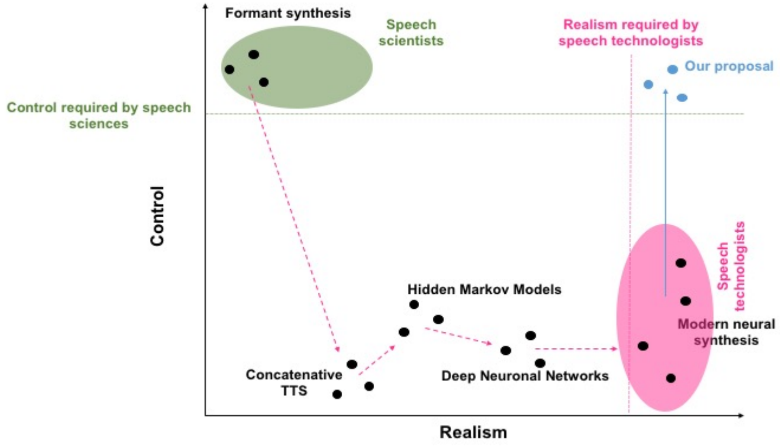

Figure 1. Schematic history of speech synthesis: black dots are TTS systems, dashed arrows show technological paradigm shifts. The direction for the future suggested in our proposal in blue.

ing controllable speech synthesis interfaces. Apart from stimulus creation, controllable speech synthesis is also able to offer whole frameworks used for testing phonological models (analysis by synthesis, e.g. Cerňak et al., 2017; Xu and Prom-on, 2014).

TTS systems that offer low-level feature control include rule-based formant synthesis. These legacy systems, however, generate a signal with impoverished perceptual cues that has low intelligibility and was proven to overburden attention and cognitive mechanisms resulting in slower processing times (Winters and Pisoni, 2004). Therefore, the validity of use of formant synthesis in e.g. stimuli creation is greatly restricted due to the many differences in perception of natural and classical synthetic speech. In essence, what these systems compensate with controllability, they lack in realism and intelligibility.

Concatenative signal generation methods, although dominant in TTS applications until recently due to their superior quality, were largely excluded from use in phonetics, as they are not able to provide a continuum of acoustic cues in response to input control. One notable exception is MBROLA (Dutoit et al. 1996), which uses a waveform-modification technique similar to PSOLA (Moulines and Charpentier, 1990) to allow control of pitch and duration given a sequence of allophones to speak.

Modern systems, that is, statistical-parametric and neural sequence-to-sequence systems are able to control arbitrary concepts by learning mappings via supervised machine learning. These concepts are usually hard to define acoustically (speaker identity, age, and gender (Luong et al., 2017), emotional state (Henter et al., 2018) and prosodic prominence (Malisz et al., 2017). So far, controllability of low-level acoustic parameters, essential for phonetic research, have not attracted the attention of speech engineers, whose systems typically are developed for commercial applications.

It thus appears that it will be up to publicly funded academic institutions and a renewed dialogue between speech researchers and speech engineers to take up this task. Therefore, we put forward our proposal concerning the steps needed towards re-connecting the goals and methods of speech sciences and technologies invested in speech synthesis, in a manner suggested in the top right corner of Fig. 1.

Our proposal

As summarised in Fig. 1, speech science and speech technology so far have been pursuing orthogonal goals. Control and realism have to be brought back into dialogue again in order for both fields to benefit.

First of all, phonetics needs speech synthesis systems that sound as close to natural speech to remove the problems listed in (Winters and Pisoni, 2004). We propose to start with validating the current achievements and demonstrating how close we actually are to generating synthetic speech that is indistinguishable from natural speech on relevant perceptual measures. What has so far been lacking is a comprehensive evaluation programme of state-of-the-art systems using precise and robust measures. It is important that the evaluation methods employed stand up to scrutiny of both the technology and research communities.

Taking this as guidance, in our recent study (Malisz et al., 2019), we showed that modern systems are substantially closer to natural speech than formant synthesis, according to a rigorous naturalness rating measure. Reaction times for several modern systems in the same study also did not differ substantially from natural speech, meaning that the processing gap observed in older systems is no longer evident. Importantly, some speech-to-speech methods were nearly indistinguishable from natural speech on both naturalness and processing measures.

Secondly, phonetic research needs controllable speech synthesis in order to fulfil its mission to a) disentangle and comprehend the perceptual role of different types of information in speech signals and b) to generate entirely new lines of research into speech phenomena that cannot be easily elicited or controlled in the lab.

Regarding a), we need to develop techniques for controlling speech- generating systems beyond what is currently possible. Our strategy envisages the use of modern technologies that prove to offer realism on the level benchmarked by studies such as Malisz et al. (2019).

For example, a controllable neural vocoder is an option in which currently used low-level acoustic parameters (such as MFCCs as shown in Juvela et al., 2018) are replaced with more phonetically meaningful speech parameters such as formant frequencies or phonological features. These same parameters can also be predicted from text and/or allophone sequences with the use of controllable end-to-end systems such as Tacotron (Wang et al., 2017). Control of prominence or other high-level features can be added to this system, as demonstrated with statistical-parametric methods in e.g. Malisz et al. (2017). With enough resources, the proposed paradigm might surpass the realism attained by PSOLA when manipulating pitch and duration.

In connection to b), we envisage that the improved systems are going to generate new areas of research. For example, speech synthesisers are now capable of generating conversational phenomena such as hesitations, backchannels, breaths, and/or non-phonemic clicks, e.g. by extending the successful, token-based approach in Szekely et al. (submitted). As natural examples of such phenomena are difficult to elicit from human speakers in empirical designs, the ability to synthesise these phenomena on demand would greatly benefit their systematic study.

Conclusion

History shows that the pursuit of realism and controllability benefits both speech sciences and speech technology. Phonetic sciences, in particular, stand to gain deeper insights from more ecologically-valid synthetic speech stimuli as well as entirely new lines of research. In order to achieve this, we can use and adapt modern speech synthesis systems that have already reached levels of naturalness comparable to natural speech. Additionally, with this contribution, we would like to signal that input from the phonetic research communities to identify suitable research targets is needed.

Acknowledgements

ZM and GEH thank Jens Edlund for helpful discussions. Grant support for ZM, JB: Swedish Research Council no. 2017-02861; GEH: Swedish Foundation for Strategic Research no. RIT15-0107; CVB and OW: EPSRC Standard Research Grant EP/P011586/1; JG: Swedish Research Council no. 2013-4935.

References

Cerňak, M., Beňuš, Š., & Lazaridis, A. (2017). Speech vocoding for laboratory phonology. Comput. Speech Lang. 42, 100–121

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & Van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for noncommercial purposes. Proc. ICSLP 1393–1396.

Henter, G. E., Lorenzo-Trueba, J., Wang, X., & Yamagishi, J. (2018). Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. arXiv preprint arXiv:1807.11470.

Juvela, L., Bollepalli, B., Wang, X., Kameoka, H., Airaksinen, M., Yamagishi, J., & Alku, P. (2018, April). Speech waveform synthesis from MFCC sequences with generative adversarial networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5679-5683). IEEE.

King, S. (2015). What speech synthesis can do for you (and what you can do for speech synthesis). Proc. ICPhS.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition 21(1), 1–36.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. Proc. ICPhS 563–567.

Luong, H.-T., Takaki, S., Henter, G. E., & Yamagishi, J. (2017). Adapting and controlling DNN-based speech synthesis using input codes. Proc. ICASSP 4905–4909.

Malisz, Z., Henter, G.E., Valentini-Botinhao, C., Watts, O., Beskow, J., & Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: a discussion and an evaluation," in Proc. ICPhS.

Malisz, Z., Berthelsen, H., Beskow, J., & Gustafson, J. (2017). Controlling prominence realisation in parametric DNN-based speech synthesis. Proc. Interspeech 1079–1083.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9(5-6), 453–467.

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., et al., (2018). Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. Proc. ICML 4693–4702.

Székely, É., Henter, G.E., Beskow, J., & Gustafson, J. (submitted). Spontaneous conversational speech synthesis from found data. Submitted to Interspeech.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., et al., (2017). Tacotron: Towards end-to-end speech synthesis. Proc. Interspeech 4006–4010.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., et al., (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. Proc. ICML 5180–5189.

Watts, O., Henter, G. E., Merritt, T., Wu, Z., & King, S. (2016). From HMMs to DNNs: where do the improvements come from? Proc. ICASSP 5505–5509.

Winters, S. J., & Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. Research on Spoken Language Processing Progress Report (26), 95–138.

Xu, Y., & Prom-On, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. Speech Commun. 57, 181–208.