# FDT3317 Speech Synthesis from Beginning to End-to-end 7.5 credits

### Talsyntes från startpunkt till ändpunkt-till-ändpunkt

This is a translation of the Swedish, legally binding, course syllabus.

If the course is discontinued, students may request to be examined during the following two academic years

## Establishment

Course syllabus for FDT3317 valid from Autumn 2019

## Grading scale

P, F

## Education cycle

Third cycle

## Specific prerequisites

Admitted to a doctoral education programme.

## Language of instruction

The language of instruction is specified in the course offering information in the course catalogue.

# Intended learning outcomes

After having completed the course, the students should be able to:

1. Demonstrate a solid knowledge basis for doing independent research and development of state-of-the-art text-to-speech synthesis.

2. Define and motivate basic concepts in TTS-relevant acoustic phonetics and signal processing, and describe all parts of the text-to-speech pipeline.

3. Using the above understanding as a basis, acquire and demonstrate skills in system implementation, as practiced and evaluated during exercise sessions.

4. Demonstrate good familiarity with the seminal advances in speech synthesis over the years (both at KTH and at large), as well as with the most recent achievements such as neural-network-based end-to-end systems.

# Course contents

“Machines that speak” is an age-old topic that has experienced a recent surge in research interest. Speaking devices are now in everyone's pockets, and the speech-synthesis field has become a challenging proving ground for new methods in machine learning.

This course is an introduction to text-to-speech (TTS) synthesis with elements of acoustic phonetics and signal processing. The course introduces a universal TTS engineering pipeline step by step: text processing, prediction engine, and waveform generation. The pipeline components are then explored within each contemporary speech-synthesis paradigm, from unit selection via statistical-parametric and hybrid synthesisers to end-to-end systems.

# Disposition

The course begins with a series of lectures, followed by a larger block of reading of seminal engineering and scientific work in the field. Both lecture and reading blocks will be supported with hands-on exercises performed in class and at home. We envisage a weekly workload of 3 hours of in-class study and 3 to 4 hours of independent study for 10 weeks.

**A. Lecture block**

1. Welcome

Practicalities; introduction to the problem; forms of communication; history and tour of synthesis at TMH

2. Signal processing

Analogue signals; digital signals; filtering; source-filter model of speech production; spectrograms; acoustic features; pitch estimation

Exercise: Filtering and pitch estimation

3. Acoustic phonetics

Articulatory phonetics and speech production; acoustic phonetics; segmental phonetics and phonology; prosodic modelling

Exercise: Phonetic transcription; spectrogram reading

4. The text-to-speech pipeline

Text analysis; dynamic time warping; hidden Markov models; decision trees; neural networks; statistical TTS paradigms

Exercise: Festival and labelling; forced alignment

**B. Discussion block**

5. Review papers on unit selection and statistical speech synthesis [1,2,3]

Exercise: Building a unit selection voice

6. Contemporary statistical parametric speech synthesis [4,5,6]

Exercise: Building a DNN-based statistical parametric synthesis voice

7. End-to-end systems [7,8,9,10]

Exercise: Building and tweaking a Tacotron or VoiceLoop end-to-end synthesiser

8. TTS evaluation [11,12]

Exercise: Building and tweaking a Tacotron or VoiceLoop end-to-end synthesiser

9. Focus session, e.g., [13,14,15]

Exercise: Evaluating the synthesisers

10. Student demonstrations

# Course literature

Suggested reading:

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. ICASSP, 1996.

[2] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Commun., 2009.

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” Proc. IEEE, 2013.

[4] Y. Qian, F. K. Soong, and Z.-J. Yan, “A unified trajectory tiling approach to high quality speech rendering,” IEEE T. Audio Speech, 2013.

[5] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, and L. Deng, “Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends,” IEEE Signal Proc. Mag., 2015.

[6] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, “From HMMs to DNNs: where do the improvements come from?,” in Proc. ICASSP, 2016.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[8] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in Proc. NIPS, 2015.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. ICASSP, 2018.

[10] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, “VoiceLoop: Voice fitting and synthesis via a phonological loop,” in Proc. ICLR, 2018.

[11] S. King, “Measuring a decade of progress in text-to-speech,” Loquens, 2014.

[12] S. J. Winters and D. B. Pisoni, “Perception and comprehension of synthetic speech,” Research on Spoken Language Processing Progress Report, 2004.

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” IEEE T. Audio Speech., 2009.

[14] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in Proc. SSW, 2004.

[15] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, “Normalization of non-standard words,” Comput. Speech Lang., 2001.

## Examination

- EXA1 - Exam, 7.5 credits, grading scale: P, F

Based on recommendation from KTH’s coordinator for disabilities, the examiner will decide how to adapt an examination for students with documented disability.

The examiner may apply another examination format when re-examining individual students.

Several components contribute to the final grade including introduction of a discussion paper, exercise participation, and final student group work on system demonstrations.

## Other requirements for final grade

A pass on all components (as listed above) is required to pass the course.

## Ethical approach

- All members of a group are responsible for the group's work.
- In any assessment, every student shall honestly disclose any help received and sources used.
- In an oral assessment, every student shall be able to present and answer questions about the entire assignment and solution.