



FDT3317 Talsyntes från startpunkt till ändpunkt-till-ändpunkt 7,5 hp

Speech Synthesis from Beginning to End-to-end

Fastställande

Kursplan för FDT3317 gäller från och med HT19

Betygsskala

P, F

Utbildningsnivå

Forskarnivå

Särskild behörighet

Antagen till forskarutbildning.

Undervisningsspråk

Undervisningsspråk anges i kurstillfällesinformationen i kurs- och programkatalogen.

Lärandemål

Efter avslutad kurs ska studenterna kunna:

1. Uppvisa en stadig kunskapsgrund för oberoende forskning och utveckling inom samtida text-till-tal-teknik.
2. Definiera och motivera grundläggande begrepp från akustisk fonetik och signalbehandling samt alla delsteg i databehandlingen från text till tal.
3. Med utgångspunkt i ovanstående teoretiska förståelse, erhålla och uppvisa kunskaper i systemimplementation, som praktiseras och uppvisas under övningar.
4. Uppvisa god bekantskap med de viktigaste framstegen i talsyntes genom åren (både vid och utanför KTH), såväl som med de senaste genombrotten såsom ändpunkt-till-ändpunktssystem baserade på djupa neurala nätverk.

Kursinnehåll

”Maskiner som kan prata” är ett klassiskt forskningsområde som på senare tid fått ökad uppmärksamhet. Idag finns talande apparater i allas fickor och talsyntes har blivit en utmanande testplats för nya maskininlärningsmetoder.

Denna kurs utgör en introduktion till text-till-tal (TTT) med delmoment i akustisk fonetik och signalbehandling. Kursen beskriver ett universellt databehandlingsflöde för TTT: textbehandlare, prediktor och vågformsgenerator. Dessa komponenter studeras sedan inom ramen för de olika talsyntesparadigmerna, från konkatenativ syntes via statistisk parametrisk syntes och hybridsyntes till ändpunkt-till-ändpunktssystem.

Kursupplägg

Kursen börjar med en föreläsningsserie. Den följs av ett större läsblock där studenterna läser framstående tekniska och vetenskapliga publikationer inom området. Både föreläsningar och läsblock kompletteras med praktiska övningar i och utanför klassrummet. Vi föreställer oss en arbetsbörda om 3 timmar lektioner och 3 till 4 timmar enskilda studier per vecka i 10 veckor.

A. Föreläsningsblock

1. Inledning

Praktikaliteter; problembeskrivning; olika former av kommunikation; historia och guidad tur av talsyntes vid TMH

2. Signalbehandling

Analog signaler; digitala signaler; filter; källa-filtermodellen; spektrogram; akustiska särdrag; tonhöjdsuppskattning

Övning: Filtrering och tonhöjdsuppskattning

3. Akustisk fonetik

Artikulationsfonetik och talproduktion; akustisk fonetik; fonetik och fonologi på segmentnivå; modeller av satsmelodi

Övning: Fonetisk transkription; spektrogramavläsning

4. Dataflöde från text till tal

Textanalys; dynamisk tidsförskjutning; dolda Markovmodeller; beslutsträd; neurala nätverk; statistisk talsyntes

Övning: Festival och annotering; tvingad matchning

B. Diskussionsblock

5. Översiktsartiklar inom konkatentativ syntes och statistisk talsyntes [1,2,3]

Övning: Bygg ett konkatentativt talsyntessystem

6. Nutida statistisk parametrisk talsyntes [4,5,6]

Övning: Bygg ett DNN-baserat statistiskt parametriskt talsyntessystem

7. Ändpunkt-till-ändpunktssystem [7,8,9,10]

Övning: Bygg och finjustera ett Tacotron- eller VoiceLoop-system för ändpunkt-till-ändpunktssyntes

8. Utvärdering av text-till-talteknik [11,12]

Övning: Bygg och finjustera ett Tacotron- eller VoiceLoop-system för ändpunkt-till-ändpunktssyntes

9. Fördjupning, t.ex. [13,14,15]

Övning: Utvärdera talsyntessystemen

10. Studentdemonstrationer

Kurslitteratur

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 - Tentamen, 7,5 hp, betygsskala: P, F

Examinator beslutar, baserat på rekommendation från KTH:s handläggare av stöd till studenter med funktionsnedsättning, om eventuell anpassad examination för studenter med dokumenterad, varaktig funktionsnedsättning.

Examinator får medge annan examinationsform vid omexamination av enstaka studenter.

När kurs inte längre ges har student möjlighet att examineras under ytterligare två läsår.

Flera moment bidrar till slutbetyget, inklusive att framlägga en diskussionsartikel, delta i övningar samt det avslutande grupparbetet med systemdemonstrationer.

Övriga krav för slutbetyg

Godkänt omdöme i alla kurskomponenter (angivna ovan) krävs för att bli godkänd i kursen.

Etiskt förhållningssätt

- Vid grupparbete har alla i gruppen ansvar för gruppens arbete.
- Vid examination ska varje student ärligt redovisa hjälp som erhållits och källor som använts.
- Vid muntlig examination ska varje student kunna redogöra för hela uppgiften och hela lösningen.