Sinteza vorbirii (TTS)

6.1 Introducere

Sinteza vorbirii (Text-to-Speech – TTS) este procesul prin care un text scris este convertit într-un semnal audio care imită vorbirea umană. Deși a fost o provocare tehnologică majoră timp de decenii, în ultimii ani, dezvoltările în rețele neuronale profunde și învățarea automată au transformat complet acest domeniu. Aplicațiile sintezei vorbirii se extind de la asistenți vocali (ex. Siri, Alexa), la cititoare de ecran pentru persoane cu deficiențe de vedere, traducere automată, educație, divertisment și interfețe conversaționale.

Acest capitol oferă o privire de ansamblu structurată asupra evoluției sintezei vorbirii, începând cu abordările clasice și culminând cu modelele neuronale end-to-end de ultimă generație. De asemenea, discutăm vocoderele neuronale, seturile de date, evaluarea sistemelor și direcțiile emergente în cercetare.

6.2 Abordări Clasice: Sinteza Concatenativă și Parametrică

Înainte de epoca rețelelor neuronale adânci, sinteza vorbirii era dominată de două paradigme fundamentale: **sinteza bazată pe concatenare** și **sinteza parametrică statistică**. Aceste metode, deși în prezent considerate învechite, au pus bazele conceptuale ale arhitecturilor moderne și încă oferă perspective utile asupra procesului de generare a vorbirii.

6.2.1. Sinteza bazată pe concatenare

Sinteza prin concatenare, cunoscută și sub denumirea de sinteză cu selecție de unități (unit selection synthesis), constă în construirea semnalului vocal sintetic prin alăturarea segmentelor de vorbire preînregistrate, extrase dintr-un corpus vorbit. Aceste segmente pot fi de diverse dimensiuni și granularități, incluzând foneme, difoni, silabe sau chiar unități prosodice mai largi. Dintre acestea, difonii — fragmentele care conțin tranziția acustică de la sfârșitul unui fonem la începutul următorului — au fost adesea preferați, întrucât permit o redare mai fluidă a coarticulației naturale a limbajului (Hunt & Black, 1996).

Procesul de sinteză presupune două etape majore: selecția unităților optime din baza de date și concatenarea lor acustică. Selecția este realizată cu ajutorul unei funcții de cost ce combină criterii de adecvare fonetică, continuitate acustică și coerență prosodică. Această funcție optimizează alegerea celor mai potrivite unități pentru contextul fonetic dat, asigurând în același timp tranziții line între ele. Astfel, sinteza obținută este guvernată de o dublă constrângere: fidelitatea lingvistică față de textul sursă și naturalețea acustică a vorbirii generate.

Sistemele de sinteză prin concatenare, precum Festival Speech Synthesis System, au reușit să producă o voce sintetică de o calitate naturală notabilă, în special în condiții de vorbire neutră și pentru enunțuri frecvente sau preînregistrate. Cu toate acestea, această metodă este puternic dependentă de dimensiunea și diversitatea corpusului. În absența unei acoperiri fonetice și prosodice suficiente, sistemul poate produce articulații nefirești sau discontinuități perceptibile în semnalul vocal. De asemenea, adaptarea stilistică — cum ar fi schimbarea accentului, adăugarea expresivității sau

controlul emoțional — este extrem de limitată, întrucât unitățile selectate nu pot fi manipulate acustic fără a deteriora calitatea sunetului.

Importanța istorică a sintezei prin concatenare constă în faptul că a stabilit un standard ridicat al calității perceptuale, deși cu un cost ridicat în termeni de stocare și complexitate de selecție. Această metodă a fost considerată pentru mult timp „gold standard-ul" sintezei, fiind utilizată în sisteme comerciale, în navigație vocală, cititoare de e-mail și dispozitive de asistență vocală.

6.2.2. Sinteza parametrică statistică

În contrast cu rigiditatea metodei bazate pe concatenare, sinteza parametrică statistică — în special sinteza bazată pe modele Markov ascunse (HMM-TTS) — oferă un cadru flexibil, care permite o mai mare controlabilitate a vocii sintetizate. Această abordare presupune modelarea parametrilor acustici ai semnalului de vorbire prin funcții de densitate statistică, învățate dintr-un corpus antrenat, și generarea acestora pornind de la o reprezentare lingvistică derivată din textul sursă (Tokuda et al., 2000; Zen et al., 2009.

Funcționarea unui sistem HMM-TTS presupune parcurgerea unui lanț de etape interdependente. Inițial, textul introdus este transformat într-un set de trăsături lingvistice care includ informații fonetice, sintactice și prosodice, precum fonemele individuale, poziția acestora în cuvânt și propoziție, accentul, pauzele sintactice sau etichete de intonație. Aceste trăsături sunt apoi utilizate pentru a ghida selecția modelelor HMM corespunzătoare, care descriu distribuții probabilistice ale caracteristicilor acustice: spectrul (reprezentat adesea prin coeficienți cepstrali mel-frecvență – MFCC sau MGC), înălțimea tonală (pitch, F_0), și energia.

Pentru fiecare unitate lingvistică (de regulă un fonem), este antrenat un model HMM, cu stări ascunse care reflectă variațiile interne ale acestuia. La sinteză, pe baza succesiunii de etichete lingvistice, sistemul selectează secvența optimă de modele și generează media distribuțiilor acustice asociate. Acest pas este urmat de reconstruirea semnalului audio printr-un vocoder parametric — de exemplu, STRAIGHT sau WORLD — care transformă parametrii numerici în undă sonoră.

Un avantaj major al acestei metode îl reprezintă posibilitatea de a controla detaliat vocea generată, inclusiv timbrul, vorbitorul, stilul de vorbire sau expresivitatea, prin ajustarea parametrilor de intrare sau prin antrenarea pe seturi de date specializate. În plus, sinteza HMM este relativ robustă la lipsa completitudinii corpusului, fiind capabilă să generalizeze pe baze statistice în contexte noi.

Totuși, sinteza parametrică suferă de limitări perceptuale importante. Calitatea semnalului generat este afectată de tendința modelelor statistice de a supraînmedia valorile parametrilor, ceea ce duce la o voce percepută ca „plată", „robotică" sau lipsită de micro-variațiile naturale ale vorbirii umane. Aceste efecte sunt cunoscute sub denumirea de „problemă a generării prin medie" și constituie una dintre principalele critici aduse metodei HMM-TTS. De asemenea, vocoderele parametrice introduc frecvent artefacte spectrale, ce afectează fidelitatea percepută.

În pofida acestor limitări, sinteza parametrică a constituit un pas esențial în tranziția către modelele neurale end-to-end, oferind o infrastructură conceptuală și tehnică fundamentală pentru învățarea reprezentărilor acustice condiționate de contextul lingvistic.

6.3 Modele Neuronale Acustice: De la Tacotron la FastSpeech

O schimbare de paradigmă a apărut odată cu introducerea rețelelor neuronale de tip encoder-decoder cu atenție, începând cu Tacotron (Wang et al., 2017). Acesta mapează o secvență de text într-o reprezentare acustică (mel-spectrogramă) folosind un encoder bidirecțional și un decoder LSTM, asistat de un mecanism de atenție care determină „la ce parte din text să se uite" modelul la fiecare pas. Tacotron 2 (Shen et al., 2018) a îmbunătățit arhitectura originală prin utilizarea vocoderului WaveNet în locul reconstrucției Griffin-Lim, obținând o calitate a vocii comparabilă cu cea umană. Totuși, aceste modele erau lente, greu de paralelizat și sensibile la erori de aliniere.

Pentru a rezolva aceste probleme, au fost propuse arhitecturi inspirate de modelele Transformer:

- **Transformer TTS** (Li et al., 2019) a introdus encoderi și decodere complet atenționale, fără recurență.
- **FastSpeech** (Ren et al., 2019) a decuplat generarea duratei de generarea spectrogramului, folosind un *Duration Predictor*.
- **FastSpeech 2** (Ren et al., 2020) a îmbunătățit expresivitatea prin integrarea de pitch, energy și variabilitate temporală.

Aceste modele sunt rapide, stabile și pot fi antrenate eficient la scară mare.

6.4 Vocodere Neuronale: WaveNet, Parallel WaveGAN, HiFi-GAN

Ultimul pas în sinteza vorbirii constă în transformarea spectrogramelor în semnal audio real. Tradițional, această conversie era realizată cu vocodere deterministe (ex. Griffin-Lim), dar ele produceau distorsiuni auditive.

WaveNet (van den Oord et al., 2016) a fost primul vocoder neuronal, capabil să genereze semnal audio eșantion cu eșantion, într-un mod autoregresiv, condiționat de spectrogramă. Deși produce sunet de calitate înaltă, viteza de inferență este scăzută.

Pentru a permite sinteza în timp real, au apărut modele non-autoregresive, bazate pe GAN-uri:

- **Parallel WaveGAN** (Yamamoto et al., 2020) folosește discriminatori spectrali pentru a asigura realismul acustic.
- **HiFi-GAN** (Kong et al., 2020) utilizează discriminatori multipli, la mai multe rezoluții temporale și de frecvență, oferind sinteză rapidă și de înaltă fidelitate.

HiFi-GAN este acum standardul în vocoderele neurale rapide și precise, fiind utilizat în sisteme open-source precum ESPnet, Glow-TTS și Coqui TTS.

6.5 Modele End-to-End: VITS și YourTTS

Unificarea tuturor etapelor sintezei într-un singur model a fost realizată de **VITS (Variational Inference TTS)** (Kim et al., 2021), care combină un encoder textual, un model latent variațional (VAE) și un GAN pentru generarea formei de undă. VITS este antrenat end-to-end, fără pre-aliniere, învățând implicit duratele, prozodia și stilul. Calitatea sunetului este remarcabilă, iar viteza de inferență este potrivită pentru aplicații reale.

Extensii precum **YourTTS** (Casanova et al., 2022) adaugă capabilități de sinteză multi-vorbitor, zero-shot TTS (folosind doar câteva secunde de voce ca exemplu) și conversie de voce. Acestea permit generarea de voci noi, stilistice sau expresive, cu un grad ridicat de control.

6.6 Seturi de Date și Evaluarea Calității

Antrenarea și compararea sistemelor TTS necesită date de vorbire curate și diversificate. Printre cele mai utilizate corpusuri se numără:

- **LJSpeech**: o singură vorbitoare, engleză, ~24h
- **VCTK**: multi-vorbitor, engleză britanică
- **LibriTTS**: corpus derivat din audiobooks
- **Common Voice**: crowdsourced, multilingv
- **Blizzard Challenge corpora**: standard în evaluări comparative

Evaluarea se face atât obiectiv, prin metrici precum MCD, STOI, PESQ, cât și subiectiv, prin teste MOS (Mean Opinion Score), unde ascultători umani evaluează naturalețea și inteligența vorbirii.

6.7 Direcții de Cercetare și Provocări Deschise

Printre tendințele actuale se numără:

- **Sinteza zero-shot și few-shot**, pentru adaptare la voci noi cu puține date (AdaSpeech, YourTTS)
- **TTS expresiv**, condiționat pe emoție, context sau intenție
- **TTS multimodal**, integrat cu video sau gestică
- **Integrarea cu LLMs** – sinteză bazată pe înțelegerea semantică profundă
- **Optimizarea pentru embedded & mobile** – sinteză rapidă pe dispozitive cu resurse limitate

De asemenea, etica utilizării vocii sintetice, protecția identității vocale și detectarea conținutului generat devin subiecte de interes major.