



UPPSALA  
UNIVERSITET

Introduktion till språkteknologi — oktober 2008

# Talsyntes – historia och metoder

Mats Dahllöf (presentation efter Pétur Helgason)





# Text-till-talsystem — grundstenarna

Alla text-till-talsystem är datorbaserade

Text-till-talsystem har två huvudkomponenter

- Textbearbetning (text till representation av ljud)
- Syntesapparat (ljudrepresentation till ljud)

Text-till-tal sker i 3 steg

- Steg 1 — utvinna information ur texten
- Steg 2 — översätta informationen till en representation som syntesapparaten kan arbeta utifrån
- Steg 3 — syntesapparaten genererar ljud utifrån representationen



# Ljudrepresentation

FONEM: ljud som enheter i språk.

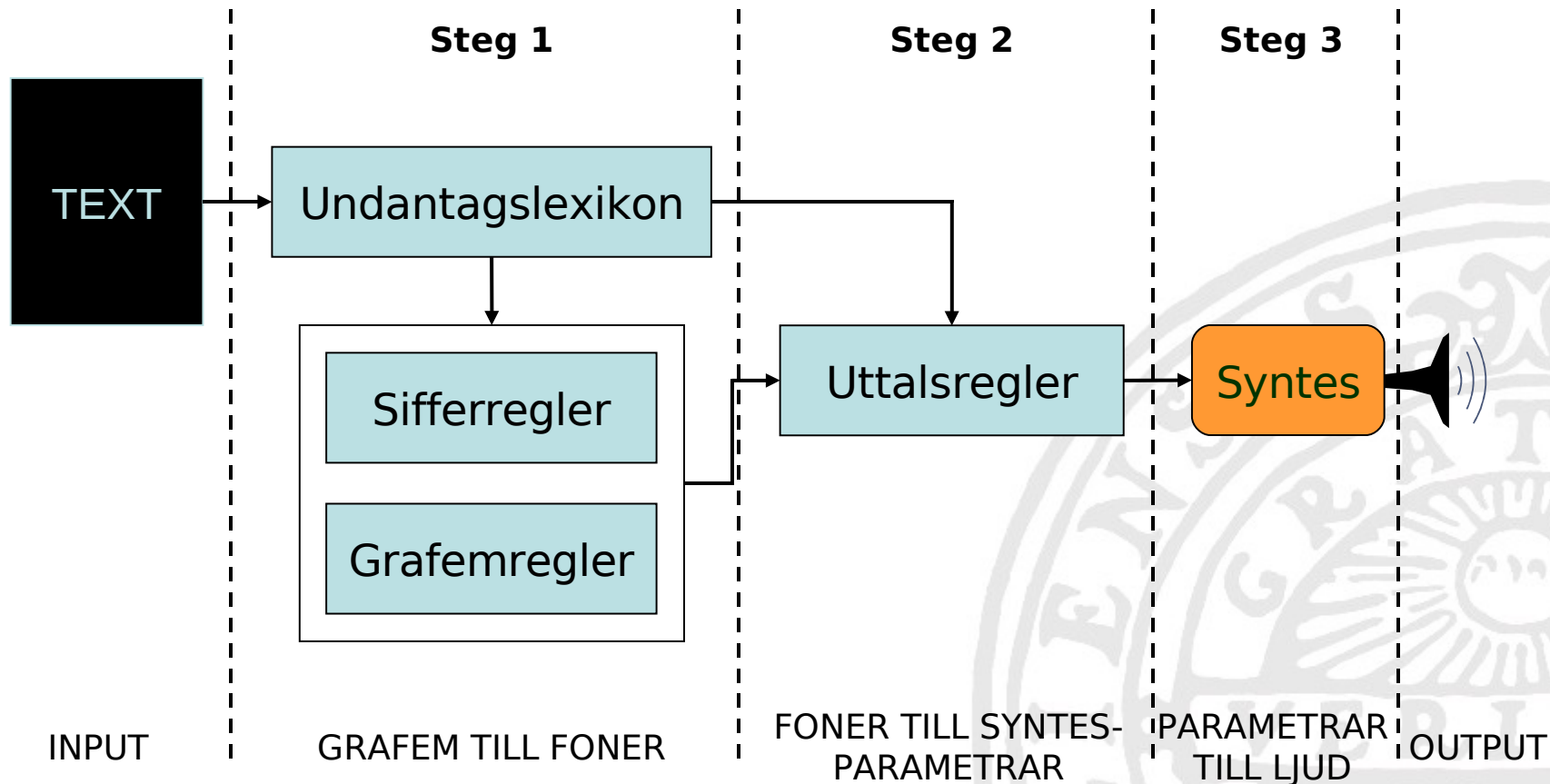
FONER: mer konkreta ljud.

T.ex. är vokalen ö ett fonem i svenskan, men den låter olika i *höra* och *hög*, om vi talar "rikssvenska". Vi kan se detta som två olika foner, mer öppet ö framför *r* och mer slutet i andra kontexter.

Annars låter det dialektalt eller fel.



# Text-till-talsystem — struktur





# Två typer av syntes

## Formantsyntes — ljudgeneratorsyntes

- Ljudgeneratoren kan alstra olika typer av ljud
  - Periodiskt ljud (röstkälla); Aspiration; Friktion
- Parametrar till formantsyntes
  - Vilken typ av ljud ska aktiveras vid en given tidpunkt
  - Vilken prosodi ska tillämpas (duration och tonhöjd)

## Kontatenativ syntes — klippa-och-klistrasyntes

- Förinspelade talsnuttar sammanfogas
- Val som måste göras:
  - Vilka snuttar ska ingå i den aktuella talsekvensen?
  - Vilken prosodi ska tillämpas? ( $F_0$  hos snuttarna kan modifieras.)



# Maskinellt tal I historien

## Uppllysningen — de stora framstegen på 1700-talet

- Ökade kunskaper om varför talet låter som det gör
  - Rösten är ett periodiskt, harmoniskt ljud
  - Röstljudet formas av talröret

## Modellering av vokaler

- Christian Kratzenstein (dansk verksam i Ryssland) gjorde fysiska vokalmodeller (1779)



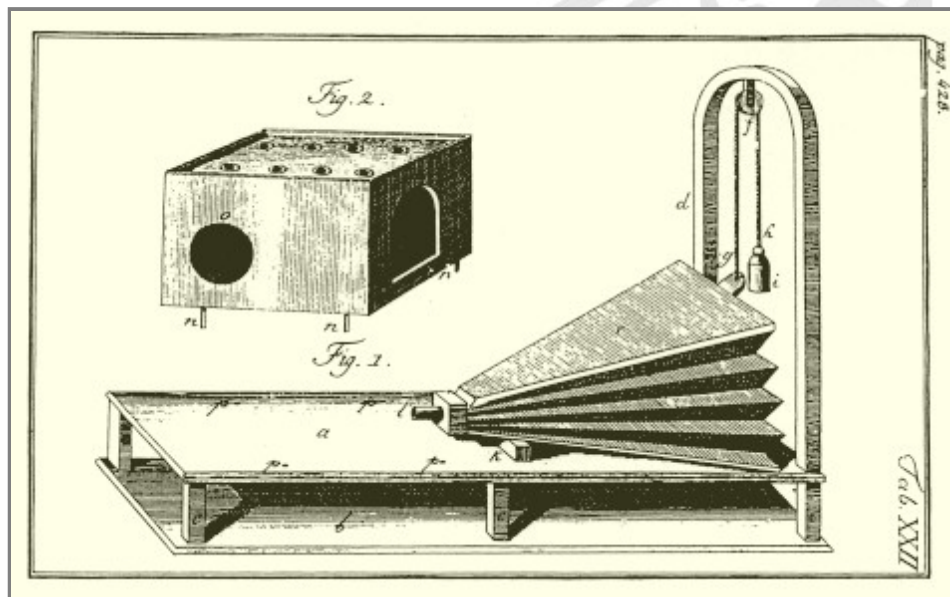
# Den första talmodelleringen

## Wolfgang von Kempelens talande maskin

- Beskrivs 1791 i skriften *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine* Wien, J. B. Degen.

## Styrmekanismer

- Lungorna simulerades med en blåsbälg
- Talet formades i en låda som dolde en rad mojänger

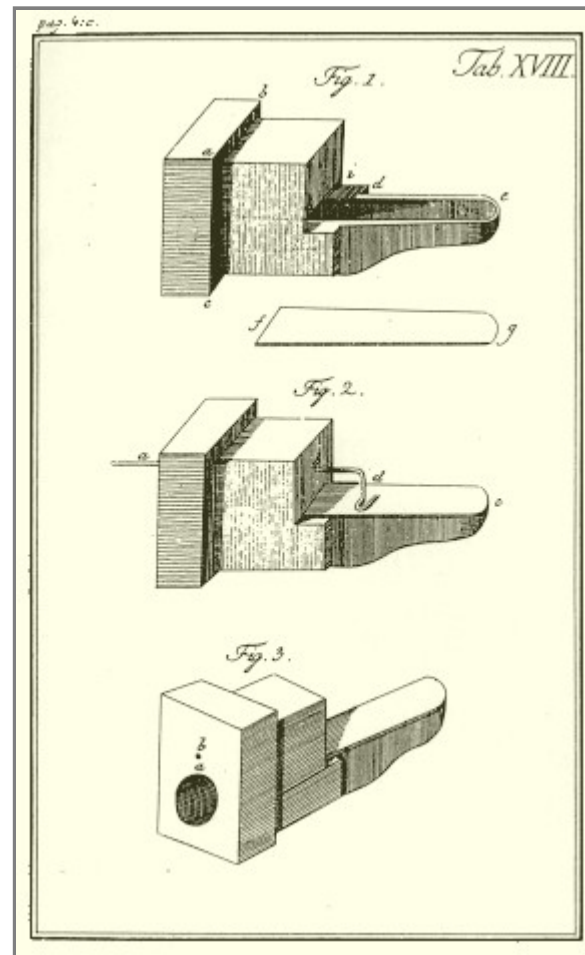




# Den första talmodelleringen

## Stämbandston

- Ett elfenbensblad i lådans inledande del simulerade stämbandston
- I en version av maskinen går det att styra längden på bladet och således ändra tonhöjden

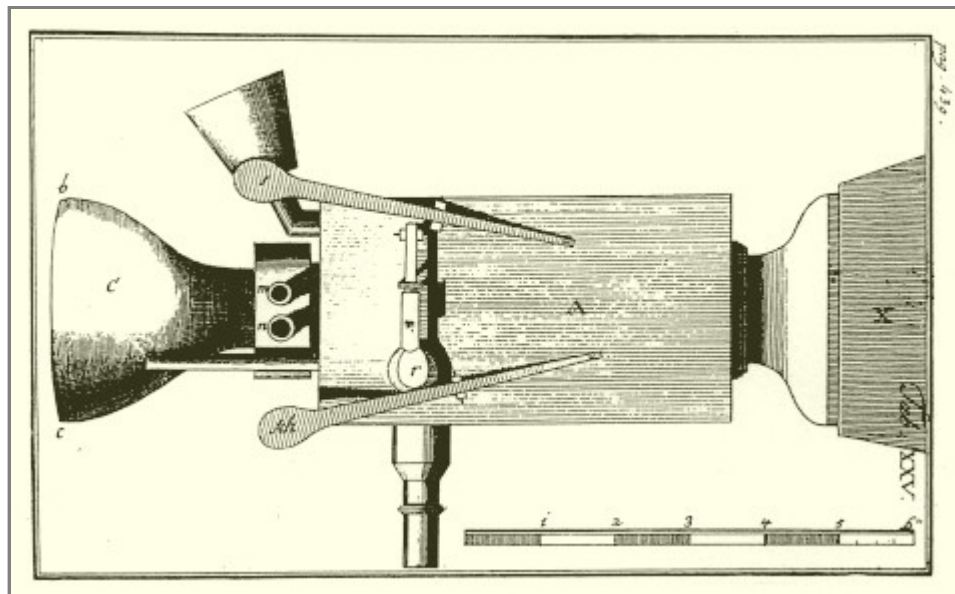




# Den första talmodelleringen

## Styrmekanismer

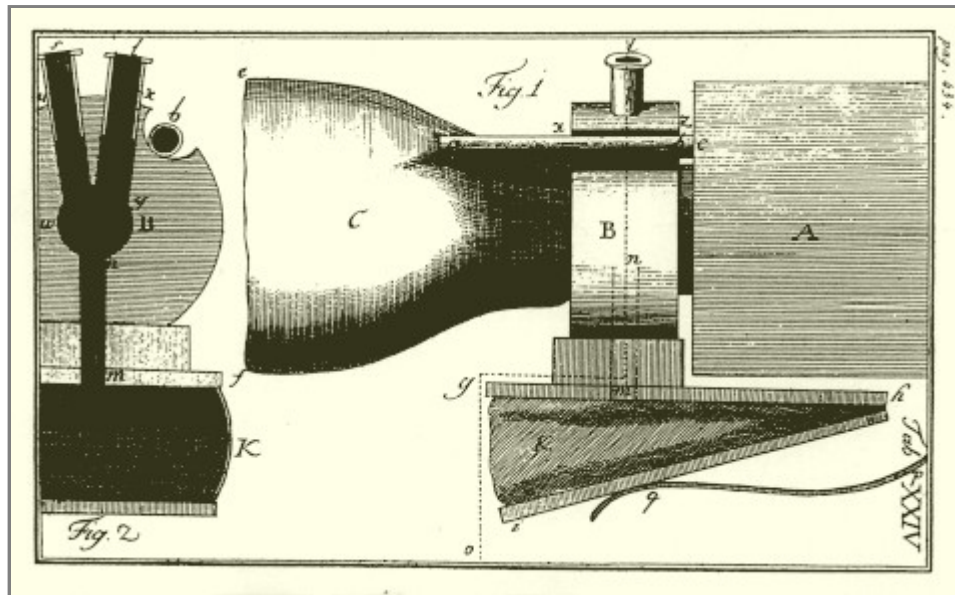
- Vokaler kunde formas genom att manipulera maskinens “mun”
- Spakar öppnade kanaler med vilka olika sibilanter (t.ex s) skapades
- Två näsborrar fanns som fick täppas till om inte en nasal skulle göras





# Den första talmodelleringen

Styrmekanismer: En extra blåsbälg under boxen användes för att skapa extra tryck för klusilproduktion



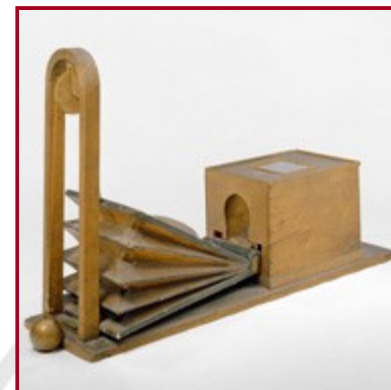
Enkelt? Von Kempelen framhöll att efter endast tre veckors träning kunde man uppnå en hyfsad kompetens i att framställa tal



# Den första talmodelleringen

## Hur lät von Kempelens tallåda?

- Vi får höra de tyska orden “es war” som först uttalas av en kvinna och sedan med Kempelens talmaskin.
- Sedan får vi höra den engelska meningen “I go” på samma sätt
- Slutligen får vi höra franskans “je t’aime”



## Fanns det någon vits med detta?

- Von Kempelens försök ökade förståelsen för vilka faktorer i artikulationen styr det akustiska resultatet



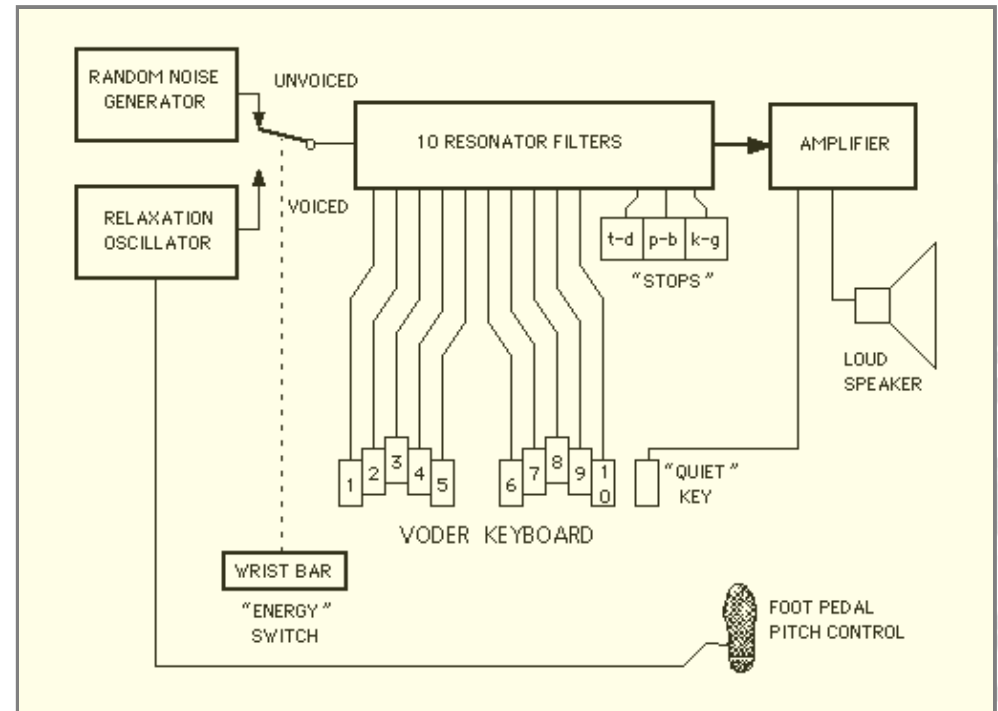
# 1939 — The Voder

## Den första moderna talsyntesapparaten

- Ljudet framställs på elektronisk väg
- Styrs med tangenter och en pedal

## Manuell styrning

- Ingen textinmatning, d.v.s. inte text-till-tal





# Voder

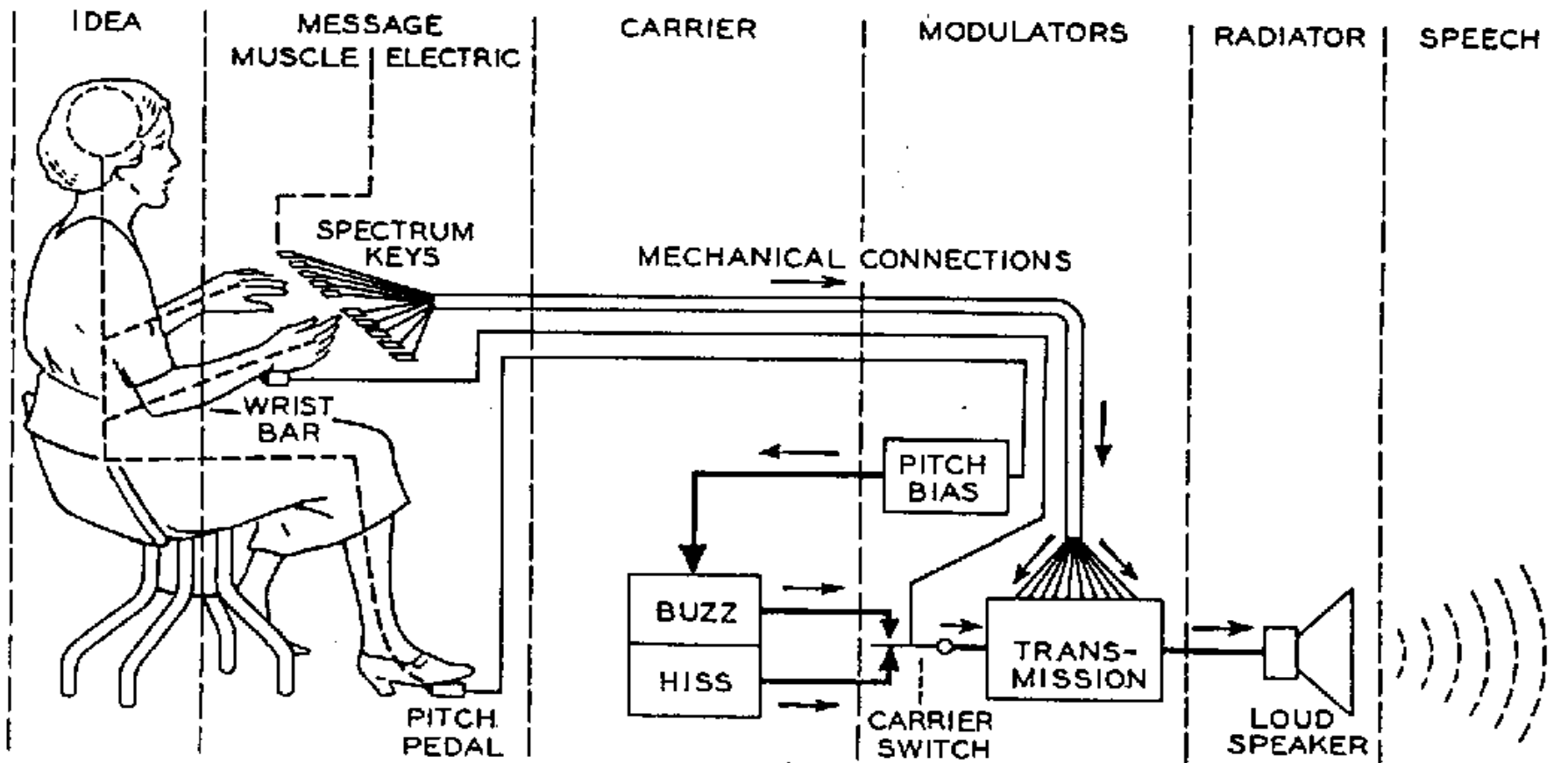
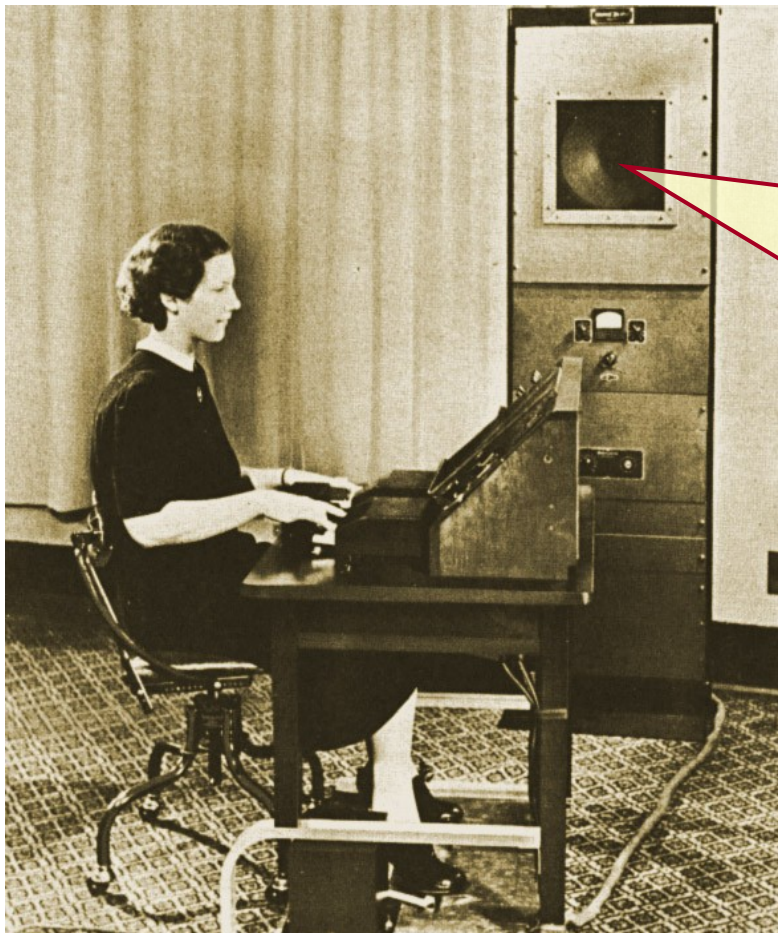


Fig. 8—Schematic circuit of the voder.



# 1939 — The Voder: hur lät den, då?



“Will you please make the Voder say for our eastern listeners: ‘Good evening, radio audience’.”

“Good evening,  
radio audience.”

“And now for our western listeners say: ‘Good afternoon, radio audience’.”

“Good afternoon,  
radio audience.”

# 1950-talet — OVE (Orator Verbis Electricis)

## OVE I

- Talsyntesapparat som utvecklades av Gunnar Fant på KTH
- Styrs med en styrstång på en tvådimensionell yta (demoprogram finns)
- Ingen textinmatning, d.v.s. inte text-till-tal
- Egentligen kunde OVE I endast göra



“How are you?” — “I love you.”



# 1961 — Första talsyntesen på dator

## Bell Labs datorsyntes

- Kelly & Gerstman på Bell Labs skapade en datorbaserad talsyntesapparat
- Den kördes på dåtidens värstingdator, en IBM 704
- Än så länge är det dock inte fråga om text-till-tal

“To be, or not to be, that is the question.  
Whether 'tis nobler in the mind to suffer  
The slings and arrows of outrageous fortune.”





# 1968 — Första text-till-talsystemet

## Text-till-tal av Noriko Umeda m fl

- Texten omvandlas först till foner
- Input till syntesapparaten är en fonisk text.
- Fonerna tolkas om till parametrar som matas in till syntesapparaten
- Syntesens kvalitet upplevs dock som mycket dålig

“Once upon a time there lived a king and queen who had no children. Not a day passed that the queen did not say: ‘If only we had a child’. One day, as the queen was walking beside the river, a little fish lifted its head out of the water.”



## Status år 1970

### Formantsyntesen kan generera acceptabelt resultat

- Man kan analysera ett stycke inspelat tal och härleda formantsyntesparametrar direkt från inspelningen
- Matar man in de siffrorna till formantsyntesapparaten blir resultatet förbluffande likt originalet

PAT 1962

OVE II 1962

OVE II 1961

PFS 1973

“Welcome to the Stockholm Speech  
Communication Seminar.”

“I enjoy the simple life.”  
“He knows just what he wants.”

“I enjoy the simple life, as long as there’s  
plenty of comfort.”



# Formantsyntesens utmaning

## Problemet

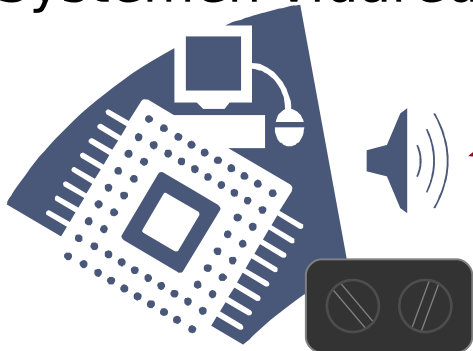
- Att automatiskt omvandla en textsträng till syntesparametrar är mycket svårare än att ställa in parametrarna direkt så att de härmar ett stycke tal



# Text-till-tal system för svenska

## Tal, musik och hörsel (TMH) på KTH — 1970–1995

- Formantsyntes (GLOVE) — Carlson och Granström
- Språkanpassbart text-till-talsystem
  - Ett "lingvistanpassat" programspråk (RULSYS)  
UA → B / C \_ D
  - Bl.a. utvecklades text-till-talsystem för svenska, norska, danska, engelska, tyska, franska och isländska.
- Systemen vidareutvecklades senare kommersiellt

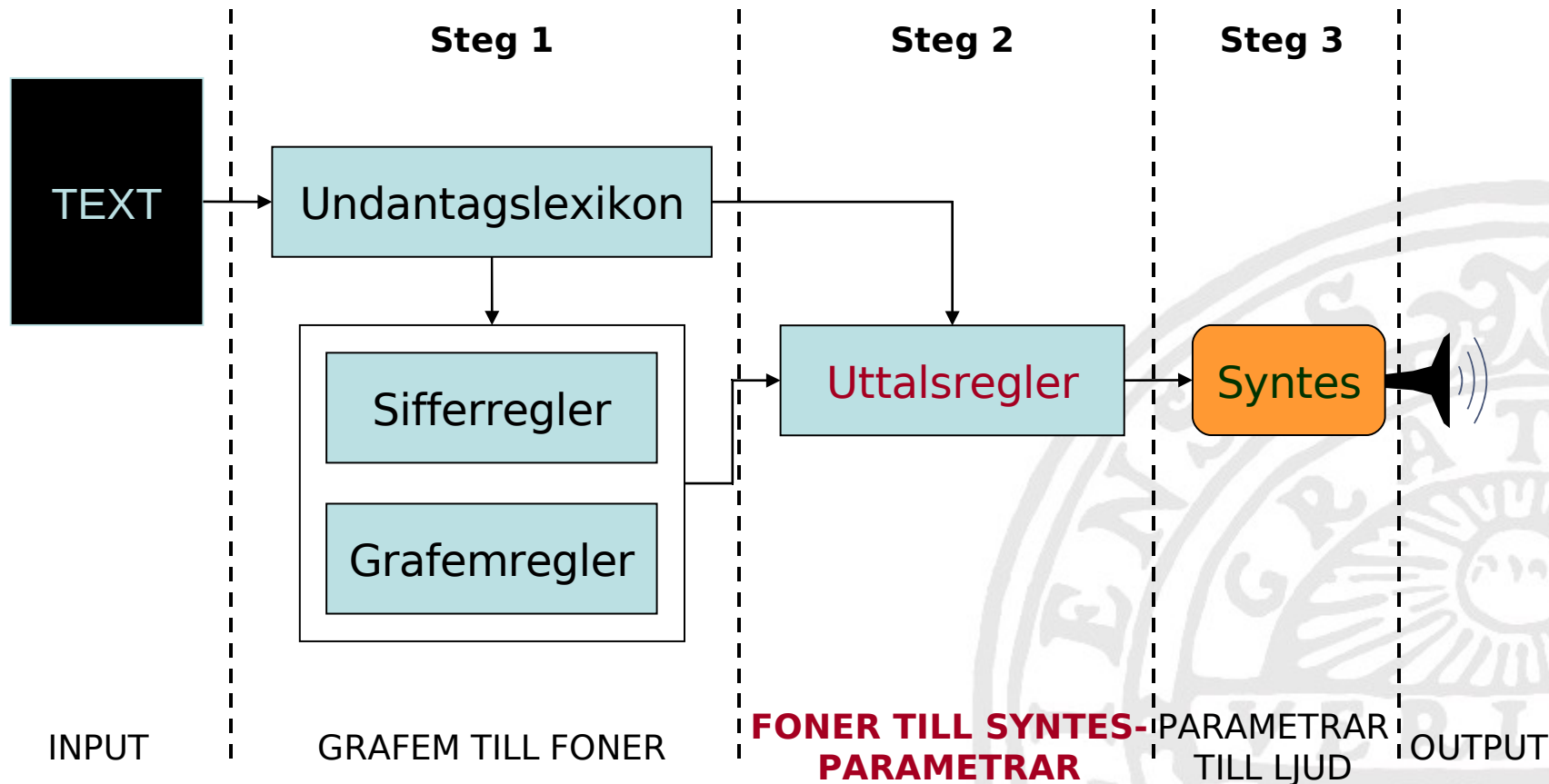


Detta är en demonstration av flerspråkigt syntetiskt tal utvecklat av Telia Promotor.

Produkten är en komplett text-till-talomvandlare som accepterar en godtycklig text, utan restriktioner avseende ordval eller meningstyp.



# Foner till formantsyntesparametrar





# Foner till formantsyntesparametrar

Varje fon har en ljudlig specifikation i systemet

- [e:] — duration; ljudstyrka; ljudkälla; formanter; formantbandbredd; formanttransitioner, etc.
- [s] — duration; ljudstyrka; ljudkälla; excitationsfrekvens, brusbandbredd, etc.

Specifikationerna måste anpassas till kontexten

- I /epra:/-delen av sekvensen *Kalle pratar* måste fonerna [e], [p], [r] och [a:] smälta ihop på rätt sätt utifrån specifikationerna för de enskilda fonerna
- Detta är en mycket svår uppgift



# Formantsyntesen passé?

## Konkatenativ syntes

- Att generera bättre formantsyntesparametrar från text har blivit allt mindre intressant p.g.a. ny syntesteknologi
- Konkatenativ syntes innebär att man spelar in en talare och gör snuttar av inspelningarna som sedan kan fogas ihop till sammanhängande tal

## Problemetets lösning

- I stället för att *härma* tal genom syntes *kopierar* man naturligt tal och gör syntes av det
- Problemet med att modellera de finare detaljerna i språkets segmentella struktur kan därmed försvinna



# Konkatenativ syntes

## Två typer

- Difonsyntes
  - Alla möjliga tvåfonems-kombinationer är representerade i systemets databas
- "Unit Selection" syntes
  - Tvåfonemskombinationer samt större talenheter (t.ex. många funktions-ordssekvenser) är representerade i systemets databas

Input: *Alla som är i stan*

1.#\_a  
2...a\_l:  
3.....l:\_a  
4.....a\_s  
5.....s\_o  
6.....o\_m  
Etc.

1.#\_a  
2...al:a  
3.....a\_s  
4.....som\_är\_i  
5.....i\_s  
6.....s\_t  
Etc.



# Difonsyntes — exempel



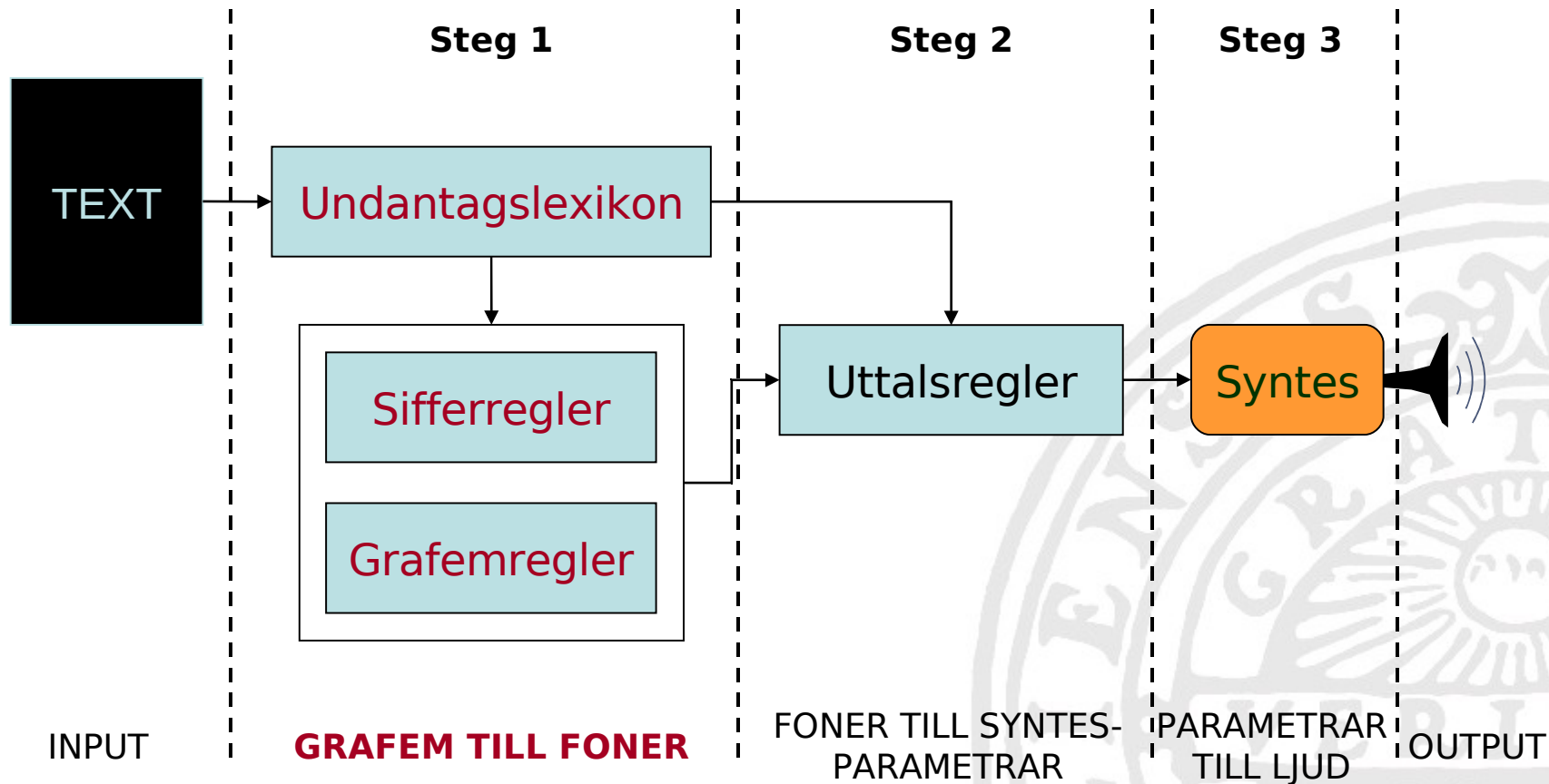
Detta är en demonstration av Infovox manliga difonsyntes utvecklat av Telia Promotor.

Produkten är en komplett text-till-talomvandlare som accepterar en godtycklig text, utan restriktioner avseende ordval eller meningstyp.

Det flexibla syntessystemet gör det möjligt att använda talad information i en mängd situationer.



# Grafem till foner





# Grafem till foner

## Grafemregler

- Regelbundna förhållanden mellan stavning och fonetisk representation kan omsättas till regler
  - /r/ + /t,d,n,l,s/ sammansmälter till retroflexa konsonanter
  - /ö/ är öppnare före r än före andra konsonanter

## Undantagslexikon

- Rymmer alla undantag från reglerna
  - *Urdu* och *Saturnus* måste finnas med i undantagslexikonet eftersom retroflexregeln inte ska tillämpas
- För svenskans del är det mycket låneord i undantagslexikonet
  - Detta eftersom betoning inte är förutsägbar i svenska
  - Grafemreglerna sätter alltid trycket på första stavelsen i ett ord, men det blir fel för de flesta låneord



# Konkatenativ syntes: svårigheter

## Kvarvarande problem

- Att lyckas bra med inspelningen av det ljudmaterial som ska ingå i syntesen är inte givet
- Eventuella problem i omvandlingen av grafem till foner kvarstår, t.ex.
  - Ordton: anden ~ anden
  - Sammansättningar: sjukanalsljudanläggning; koddatering
- Man behöver fortfarande en bra prosodisk modell
  - Tonhöjd, tryck och kvantitet kommer inte på köpet i konkatenativ syntes — de måste modelleras
  - Om man märker att någonting är fel i syntesen är det i de flesta fall något problem med prosodin